TU Clausthal

Clausthal University of Technology

ICOLE 2007, Lessach, Austria

Jacek Błażewicz, Klaus Ecker, Barbara Hammer (Eds.)

IfI Technical Report Series **IfI-07-03**

Department of Informatics
Clausthal University of Technology

Impressum

Publisher: Institut für Informatik, Technische Universität Clausthal
Julius-Albert Str. 4, 38678 Clausthal-Zellerfeld, Germany
Editor of the series: Jürgen Dix
Technical editor: Wojciech Jamroga
Contact: wjamroga@in.tu-clausthal.de
URL: http://www.in.tu-clausthal.de/forschung/technical-reports/
ISSN: 1860-8477

The IfI Review Board

ICOLE - 2007
German - Polish Workshop on Computational Biology, Scheduling and Machine Learning
Lessach, 27.05. - 02.06.2007

Jacek Błażewicz, Klaus Ecker, Barbara Hammer (Eds.)

Managing Editors: Jan Biel, Alexander Hasenfuss, Bassam Mokbel

Contents

Introduction

Jacek Blazewicz, Klaus Ecker, Barbara Hammer

The German - Polish Workshop on Computational Biology, Scheduling and Machine Learning took place in the lovely Tauern Alps resort - Lessach. It was organized in the framework of the ICOLE series of workshops. The workshop gathered together 22 scientists, Ph.D. students and Master students from the Clausthal University of Technology (Germany), Poznan University of Technology (Poznan), and Ohio University (USA). They represented the above mentioned three fields of research: Computational Biology, Machine Learning and Scheduling. Altogether there were 6 established researchers, 2 Ph.D. students and 12 M.Sc. students.

The aim of the workshop was to present recent developments in the areas of Computational Biology, Scheduling, and Machine Learning, and to exchange ideas coming from different subjects. This could give rise to new approaches, since many problems from different areas can be modeled at a certain level of abstraction by similar tools, e.g., graph theory, and then solved using a similar methodology, e.g., machine learning approaches.

The workshop started with several introductory lectures: Jacek Blazewicz and Piotr Lukasiak presented the basic description and aims of the European Project for the Antivirus Vaccine Development - Compuvac. Barbara Hammer discussed various aspects of machine learning and addressed recent results in prototype-based learning.

Later participants presented their achievements and new developments in the field, respectively, along the lines of the workshop. This included two sessions around machine learning presented by participants from Clausthal: one session about challenges and general solution methods in artificial intelligence and one about unsupervised prototype-based methods. Several interesting new research results in bioinformatics were gathered together in a couple of sessions: one session addressed the topic of RNA analysis including structure and function analysis, presented by members who are supervised by Marta Szachniuk from Poznan University; a second session centered around new results for protein analysis including, among other aspects, state-of-the-art results on domain recognition for protein structures. This was presented by members of the group lead be Piotr Lukasiak from Poznan University. Aspects of scheduling focused, on the one hand, on assembly line scheduling and car manufacturing presented by stu-

dents supervised by Grzegorz Pawlak, on the other hand on scheduling of coupled tasks, presented by Michal Tanas.

The last session combined all aspects addressed in the workshop: Klaus Ecker (Ohio University) summarized new methods for motif discovery and its combination to gene regulation; Alexander Hasenfuss (Clausthal University of Technology) presented recent machine learning techniques for general data structures such as sequences or graphs. At the end of the workshop Grzegorz Pawlak summarized the approaches used by the Poznan team in the area of scheduling in manufacturing systems.

The participants of the workshop were overwhelmed by the friendly atmosphere of the workshop and its perfect organization run by Alexander Hasenfuss. Further, Bassam Mokbel, Jan Biel, and Alexander Hasenfuss did an excellent job to transfer the heterogeneous abstracts submitted by the participants into a single LaTeX-file with a consistent layout.

Computational support for anti-virus vaccine development - COMPUVAC

***Jacek Blazewicz*[2], *Piotr Lukasiak*[1,2]**

1 Introduction

COMPUVAC: is a European collaborative project aimed at vaccine evaluation standardization and its application to the development of efficient vaccines against hepatitis C virus. CompuVac (www.compuvac.org) is a four-year integrated project (2005-2008) supported by the European Commission under its 6th Framework Programme within the "Life sciences, genomics and biotechnology for health" priority. CompuVac assembles 16 European teams from both academia and industry, in 9 European countries, with a multidisciplinary expertise in the fields of vaccine development, immunology, virology, vectorology, biomathematics and computer sciences.

2 Description

Recombinant viral vectors and virus-like particles are considered the most promising vehicles to deliver antigens in prophylactic and therapeutic vaccines against infectious diseases. Several potential vaccine designs exist but their cost-effective development cruelly lacks a standardized evaluation system. On these grounds, CompuVac is devoted to (i) rational development of a novel platform of genetic vaccines and (ii) standardization of vaccine evaluation. CompuVac assembles a platform of viral vectors and virus-like particles that are among today's most promising vaccine candidates.

CompuVac recognizes the lack of uniform means for side-by-side qualitative and quantitative vaccine evaluation and will thus standardize the evaluation of vaccine efficacy and safety by using "gold standard" tools, molecular and cellular methods in virology and immunology, and algorithms based on genomic and proteomic information. "Internal standard" algorithms for intelligent interpretation of vaccine efficacy and

[1]E-mail: Piotr.Lukasiak@cs.put.poznan.pl
[2]Institute of Computing Science, Poznan University of Technology, Poznan, Poland

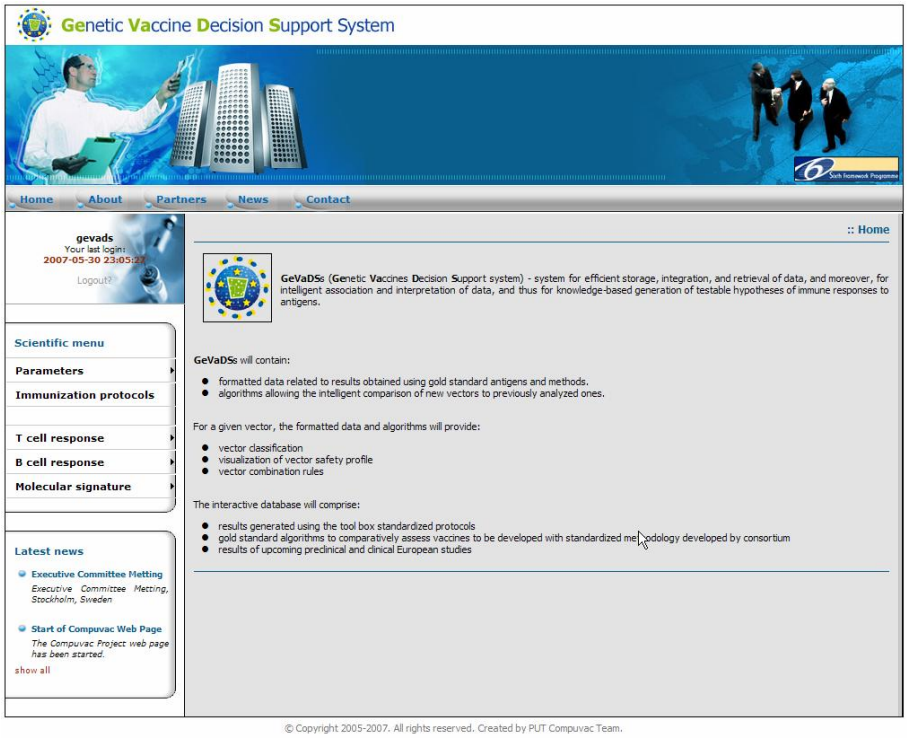

Figure 1: Screenshot of the GeVaDSs web page.

safety will be built into CompuVac's interactive "Genetic Vaccine Decision Support System", which should generate: (i) vector classification according to induced immune response quality, accounting for gender and age, (ii) vector combination counsel for prime-boost immunizations, and (iii) vector molecular signature according to genomic analysis.

The knowledge assembled from these studies will be applied to the development of vaccines. As end products, vector platform and "internal standard" tools, methods and algorithms will be available to the scientific and industrial communities as a toolbox and interactive database which standardized nature should contribute to cost-effective development of novel vaccines.

3 Summary

GeVaDSs is a unique and representative system to store, analyze and compare results obtained during vaccine development as well as all features of vaccines experiments. This system is currently available via internet http://gevads.cs.put.poznan.pl.

A new tool for fundamental analysis of NMR spectra

***Wojciech Paczkowski*[1,2], *Michal Milawski*[2], *Lukasz Popenda*[3], *Marta Szachniuk*[2,3]**

1 Introduction

Nuclear Magnetic Resonance spectroscopy is a powerful tool used to analyze structures and dynamics of biomolecules. The process of structure determination by NMR starts from performing the experiments and recording the spectra. Next, the process of studying spectral data begins. In the first analytical step spectra are examined and resonance signals are assigned to the appropriate atoms of the molecule. It is followed by a computation of structural parameters from which molecular structure is reconstructed. Each cross-peak in the spectrum represents resonance signal generated by two atoms. The procedure of signal assignment can be preceded by identification of cross-peaks, which facilitates the assignment. During identification one determines which types of atoms are responsible for generating signals recorded in the spectrum. Next particular atoms are assigned to the cross-peaks. In this way, values of chemical shifts related to the cross-peaks are determined for the selected atoms. These values carry important structural information which can be applied to derive secondary and tertiary structure of a molecule. Assignment of resonance signals is more difficult for RNA than for proteins [3]. Several tools have been developed to facilitate manual resonance assignment based on protein NMR spectra [4, 5, 6]. Since there is an observable growth of interest in RNA molecules, and nearly half of all current RNA structures have been determined by NMR spectroscopy techniques [3], it is crucial to develop automated methods for analysis of data stored in the spectra of RNAs. Here, we present a software tool designed to facilitate assignment of a collection of cross-peaks in NMR spectra obtained for RNA molecules. It can identify hydrogen atoms and related carbons resonating

[1]E-mail: wojtek.paczkowski@gmail.com
[2]Institute of Computing Science, Poznan University of Technology, Poznan, Poland
[3]Institute of Bioorganic Chemistry, Poznan, Poland

during the following two-dimensional experiments: NOESY, DQF COSY and 1H-13C HSQC. AnalyzeIT! has been written in Java language. It runs on all operating systems for which Java Runtime Environment is available.

2 Methods

AnalyzeIT! is simple to use and provides user-friendly interface. User starts from opening text files with spectral data and runs an analysis by clicking AnalyzeIT! button. The results are obtained within milliseconds and displayed in three tables (NOESY, DQF COSY and 1H-13C HSQC) which contain the proposed identification. The fundamental assignment principle of AnalyzeIT! bases on the experimentally determined ranges of proton chemical shifts. We follow the ranges given in [3] and collected from Biological Magnetic Resonance Data Bank. Aromatic and sugar protons with the related carbons can be identified by the program. Main algorithm is performed in five steps. The first one concerns an analysis of COSY-type spectrum. Resonance signals recorded in DQF COSY experiment are classified into several classes related to atom types. The results of first-step classification are compared to the information given in the NOESY spectrum of the same molecule. In the third step 1H-13C HSQC spectrum is analyzed and cross-peaks from this spectrum are recognized. Again, the results of this classification are refined on the basis of the NOESY spectrum. Finally, resonances represented in the NOESY spectrum are identified and results are presented. User can manipulate with the level of shift tolerance. This allows for a proper signal identification even if the chemical shifts of the atoms do not exactly match the defined ranges or when the same cross-peak has different coordinates within two different spectra.

3 Discussion

AnalyzeIT! has been tested on several machines, in Linux and Windows environment. The program worked effectively and quickly. The results were obtained after c.a. 0.5 - 2 ms for each tested instance (times obtained in Microsoft Windows XP environment, Genuine Intel(R) CPU, 1.67 GHz, 1.00 GB RAM). Experimental test contained NMR spectra recorded for four unpublished RNA molecules on Bruker Avance 600 MHz. In each case three types of spectra were considered. The quality of identification has been measured in relation to the manual identification done by an expert. Figure 2 presents the percentage of atoms identified in the spectra recorded for r(GCAGAGAGCG).r(CGCUCUCUGC).

4 Acknowledgments

The authors thank Mr. Karol Pasternak for supplying spectral data to the analysis. This work was partially supported by grant 3T11F00227 from the Ministry of Science and Higher Education, Poland.

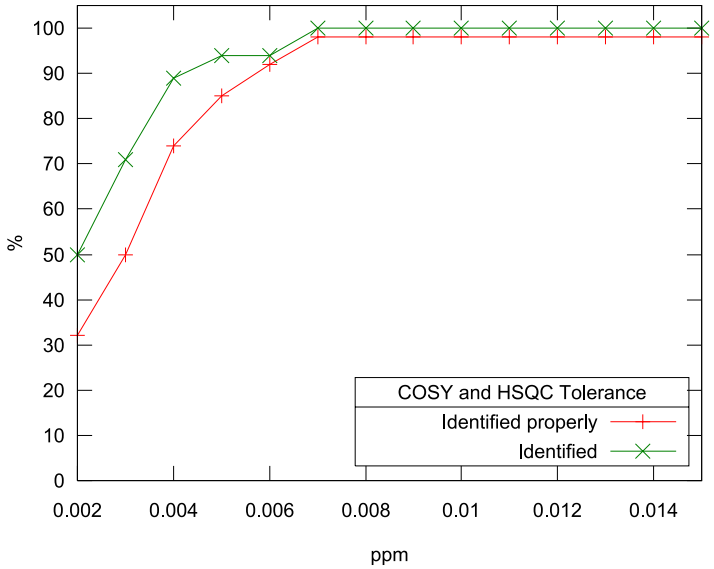

Figure 2: Percentage of properly identified atoms in exemplary molecule

References

[1] R.W. Adamiak, J. Blazewicz, P. Formanowicz, Z. Gdaniec, M. Kasprzak, M. Popenda and M. Szachniuk. An algorithm for an anutomatic NOE pathways analysis in 2D NMR spectra of RNA duplexes. *J. Comp. Biol.*, 11/1:163–180, 2004.

[2] J. Blazewicz, M. Szachniuk and A. Wojtowicz. NOE pathways construction by tabu search. *Bioinformatics.*, 21/10:2356–2361, 2005.

[3] B. Fuertig, C. Richter, J. Woehnert and H. Schwalbe. NMR spectroscopy of RNA. *ChemBioChem.*, 4:936–962, 2003.

[4] C. Mumenthaler, P. Guentert, W. Braun and K. Wuethrich. Automated combined assignment of NOESY spectra and three-dimensional protein structure determination. *J. Biomol. NMR.*, 10:351–362, 1997.

[5] W. Rieping, M. Habeck, B. Bardiaux, A. Bernard, T.E. Malliavin and M. Nilges. ARIA2: Automated NOE assignment and data integration in NMR struc-ture calculation. *Bioinformatics.*, 23/3:381–382, 2007.

[6] L. Zhang and D. Yang. SCAssign: a sparky extension for the NMR resonance assignment of aliphatic side-chains of uniformly 13C, 15N-labeled large proteins. *Bioinformatics.*, 22/22:2833–2834, 2006.

[7] Biological Magnetic Resonance Data Bank
http://www.bmrb.wisc.edu/ref_info/statsel_rna.htm.

Strategies of signal assignment in NMR spectra of RNAs with different structural motifs

Marta Szachniuk[1,2,3]***, Mariusz Popenda***[3]***, Lukasz Popenda***[3]

1 Introduction

Structural analysis of biomolecules contributes to clarify their biological functions and structure related properties in materials, components, liquid crystals, drugs, aiding identification of new diseases, raising new specimens of plants and animals, cataloging new compounds, controlling compound quality [4], etc. Initially, the research in that area concentrated on proteins and deoxyribonucleic acid (DNA). Then, it has been extended for the molecules of the ribonucleic acid (RNA), which transmits genetic information from DNA into proteins and controls certain chemical processes in the cell. There is a selection of methods of structure determination, however only two, X-ray crystallography and Nuclear Magnetic Resonance spectroscopy, give the complete information about the molecular structure at this level of study. NMR spectroscopy seems to be the best choice for probing the structure and dynamics of RNA, because of the difficulty in RNA molecules crystallization and the interest in the relationship between RNA dynamic behavior and its biological functions in solution.

Structure determination procedure using NMR is composed of two general stages: experimental, where multidimensional correlation spectra are acquired and computational, where spectra are analyzed and structure is recognized. In all methods of NMR structure analysis, raw experimental data are processed by the procedures of peak-picking, assignment, restraints determination, structure generation and refinement [3]. An assignment of the observed NMR signals to the corresponding protons and other nuclei is a bottleneck of RNA structure determination process. It is usually based on the analysis of two dimensional spectra resulting from NMR experiments and it is performed manually in accordance with the experimenter's knowledge and intuition. Since other steps of the determination process base on automatic procedures, there has been

[1]E-mail: Marta.Szachniuk@cs.put.poznan.pl
[2]Institute of Computing Science, Poznan University of Technology, Poznan, Poland
[3]Institute of Bio-organic Chemistry, Polish Academy of Sciences, Poznan, Poland

a great need to introduce automatic procedures also at this level. We have developed two algorithms for automatic generation of transfer pathways in the 2D NOESY spectra obtained for regular structures [1, 2]. Next, the methods have been modified to solve the same problem for RNA irregular duplexes. The results of the algorithms' application to the spectra recorded for RNAs with different structures have been presented.

2 Methods and Discussion

The NOE cross-peaks illustrated in the 2D NOESY spectrum are connected to form one or two paths, called transfer / NOE pathways. One pathway exists in the spectrum obtained for regular RNA duplex. Two pathways are found in the spectrum recorded for two-strand molecules with irregular structures (one pathway per a strand). Thus it is useful to know what kind of molecule is being examined before the computation starts. Considering the data specificity, automatic assignment method has been based on the combinatorial model of a NOESY graph [2]. Three graph models have been proposed to represent the problem of transfer pathway reconstruction: NOESY graph, NOESY bipartite graph and line bipartite graph. The first model has been used to model the problem and to design the algorithms. First, two heuristics, tabu search and evolutionary algorithm have been proposed for the problem of pathway reconstruction in the spectra of regular duplexes. Both have been tested on the same data coming from the real NMR experiments performed for six different RNA molecules [1]. Each solution obtained has been evaluated on the basis of its similarity to the original pathway, which has been simultaneously reconstructed by an expert. The percentage of original pathway coverage by the optimal solution has been computed. Both heuristics generated pathways covering more than 80% of the original path. Next, tabu search algorithm has been modified to suit the problem of the reconstruction of two transfer pathways, which exist in the spectra of molecules with irregular structures. The goal function has been changed and the procedure was looking for two pathways, one after the other. Solutions have been evaluated in the same way as in case of regular structures. In comparison to the spectra of regular structures, these recorded for irregular duplexes of similar size contain more cross-peaks, which is the reason of more errors made by the assignment procedure. In the spectrum containing many cross-peaks, the latter used to overlap, cover one another or join together and the algorithms have problems with differentiating between overlapping peaks. Spectra recorded for three irregular duplexes have been tested. Optimal pathways generated covered over 65% of the original solution.

3 Acknowledgments

This work was partially supported by grant 3T11F00227 from the Ministry of Science and Higher Education, Poland.

References

[1] J. Blazewicz, M. Szachniuk and A. Wojtowicz. RNA tertiary structure determination: NOE pathways construction by tabu search. *Bioinformatics.*, 21/10:2356–2361, 2005.

[2] J. Blazewicz, M. Szachniuk and A. Wojtowicz. Evolutionary approach to NOE paths assignment in RNA structure elucidation. *Proceedings of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology.*, 206–213, 2004.

[3] G. Varani and I Tinoco Jr. RNA structure and NMR spectroscopy. *Q. Rev. Biophys.*, 24:479–532, 1991.

[4] D.H. Williams and I. Fleming. RNA structure and NMR spectroscopy. *Spectroscopic Methods in Organic Chemistry. McGraw-Hill, New York*, 1996.

TU Clausthal
ICOLE 2007, LESSACH, AUSTRIA

Peptides assembling and sequencing methods

***Jacek Blazewicz*[2], *Marcin Borowski*[2], *Piotr Formanowicz*[2],**

***Tomasz Glowacki*[1,2], *Adam Kozak*[2]**

1 Introduction

The amino acid sequences of proteins determine their structure and functionality. Obviously, the genetic information encoding proteins is written in DNA. However, knowing gene sequences is not always sufficient for determining the corresponding amino acids sequences. There is no biological methods to direct prediction of the primary structure of peptides. Moreover, microscopes cannot be used because of the proteins size. In the paper the graph and theoretical models of the peptide sequence assembly have been modelled and analysed.

2 Chemical stage

In the paper the sequencing and assembling methods were defined. Sequencing is a chemical method of recognizing short peptides sequences. Assembling is a method used for finding primary structure of proteins by linking all short peptides into a long peptides' chain. Some examples of assembling applications in pharmacology and diagnostics were presented.

In our experiments, there were proposed to use two different peptidases and dividing protein material into two sets. After chemical experiment with peptidases and sequencing short peptides, one can get a context i.e. information about order of short peptides in the expected solution. We proposed Edman degradation to recognize short peptides chains. Edman degradation determines first amino acid (called N-terminal) in evaluated peptide. Iterations of Edman degradation provides information about next amino acids in searching sequence. Unfortunately, because of technological problems, one can recognize peptides up to 50 amino acids.

[1]E-mail: tglowacki@skno.cs.put.poznan.pl
[2]Institute of Computing Science, Poznan University of Technology, Poznan, Poland

3 Algorithms

The graph theory model was used to resolve assembling problem. Two graph theoretical models for peptide sequence assembling without errors have been shown and discussed. Generally, each peptide fragment which should be assembled, corresponds to a vertex in some bipartite graph. Vertices are joined by an arc with weight equal to the number of overlapping peptides in sequences corresponding to these vertices. Solving of TSP for this graph is equivalent to resolving the peptide assembly problem.

Taking into account biological experiment one can observe that there is possible to reduce the number of arcs in the graph respecting the biological dependencies. At first the arcs with weights are equal to zero may be removed from the graph, because they determinate a zero overlapping of subsequent amino acids which is infeasible. Additionally, only maximum overlapping of two small peptides is the proper one. Then the 1-graph could be constructed. As a result of the transformation the graph become the adjoint. Because of the fact that all outgoing arcs from any vertex have the same weight then TSP paths in this graph have the same value. In this specific graph looking for Hamiltonian path is equivalent to looking for Eulerian path in the other corresponding graph (see [1]). Thus, the assembling problem without errors is polynomial solvable. The computational complexity is equal to O(n+m) for proper graph representation.

Additionally, assembling problems with errors were shown and discussed. One of the sources of errors is lack of information about peptides repetitions. As a result of sequencing method one can get only information about existing peptides without information about their number. Generally, the complexity of problem with repetitions is opened. In the paper [2] polynomial time equivalent assembling problem to the exact perfect matching problem in bipartite graphs was proved. This another open combinatorial optimisation problem of unknown computational complexity. During real chemicals reaction some cuts may not occurred. The assembling problem without all cuts is strongly NP-hard which was proofed by Gallant in [3]. To introduce of this problem we proposed multi-graph representation. In this graph each vertex corresponds to the short peptide, all parallel arcs between any ordered pair of vertices determine all possible overlapping. One can take only the biggest overlapping as it is in the case of assembling without errors. The two meta-heuristics algorithms to resolve given problem have been proposed. Tabu Search [4] and Genetic algorithms [5] were presented.

4 Results

The computational experiment was performed on a set 450 instances:

- 150 random sequences with random generated peptidases
- 150 random sequences with real peptidases
- 150 real sequences with real peptidases

All experiments run 10 times for each instance. Table 1 shows an average results of these experiments:

Method	**Alignment**
Tabu Search	59,7%
Genetic Algorithm	88,74%

Table 1: Results of the computational experiment

Genetic approach gives better results than Tabu Search. The initial solution is the main reason of such big differences between derived results. There were used the effective greedy like initial solution in genetic approach and the random initial solution in Tabu Search.

5 Conclusions

In this paper assembling and sequencing methods for peptides were proposed and analysed. It has been shown that the assembling method for ideal case is polynomial time and in the case of lack of cuts the problem is NP-hard. For the second case the two meta-heuristics algorithms were constructed. The peptides assembling genetic algorithm gives much better results than Tabu Search method. In the future GRASP method will be developed because of initial solution sensitivity.

References

[1] Jacek Blazewicz, Marcin Borowski, Piotr Formanowicz, Tomasz Glowacki, "On graph theoretical models for peptide sequence assembly", Foundatons of Computing and Decisions Sciences, vol. 30 No. 3 2005, 183-191

[2] Jacek Blazewicz, Piotr Formanowicz, Marta Kasprzak, Petra Schuurman, Gerhard J. Woeginger, "DNA Sequencing, Eulerian Graphs, and the Exact Perfect Matching Problem", Lecture Notes in Computer Science 2573 (2002) 13-24

[3] Gallant J.K., "The complexity of the overlap method forsequencing biopolymers, Journal of Theoretical Biology", 101, 1983, 1-17

[4] Jacek Blazewicz, Marcin Borowski, Piotr Formanowicz, Maciej Stobiecki, "Tabu Search method for determining sequences of amino acids in long peptides" Lectures Notes in Computer Science 3449 (2005) 22-32

[5] Jacek Blazewicz, Marcin Borowski, Piotr Formanowicz, Tomasz Glowacki, "Genetic algorithm for peptide assembly problem"

Domain Analysis Server (DomAnS) – Protein domains prediction algorithm (DomAn)

Jacek Blazewicz[2,3]***, Piotr Lukasiak***[2,3]***, Maciej Antczak***[1,2]***,***

Maciej Milostan[2]***, Grzegorz Palik***[2]

1 Introduction

Protein is a large organic compounds composed of a amino acids arranged in a linear chain and joined together by peptide bonds between the carboxyl and amino groups of adjacent amino acid residues. The protein structure is much more difficult to be determined than the protein sequence (more than 4,3 million of sequence are known, but only over 43 000 3D structures are determined).

Many proteins contain compact units within the folding pattern of a single chain, that look as if it should have independent stability. These are called domains, which are globular sub-structures with more interactions within it than with the rest of the protein. Each domain containing an individual hydrophobic core built from secondary structural units connected by loop regions. The importance of domains as structural building blocks and elements of evolution has brought many automated methods for their identification and classification in proteins [12]. Automatic procedures for reliable domain assignment is essential for the generation of the domain databases. Although the boundaries of a domain can be determined by visual inspection, construction of an automated method is not straightforward.

2 Problem Formulation

There are two main definitions of problems used during domains prediction problem. In the first one . domain boundaries are predicted based on the 3D protein structure. In

[1]E-mail: mantczak@cs.put.poznan.pl
[2]Institute of Computing Science, Poznan University of Technology, Poznan, Poland
[3]Institute of Bio-organic Chemistry, Polish Academy of Sciences, Poznan, Poland

the second one domain boundaries are predicted based only on the primary structure of protein. Proposed method tries to solve the problem following second definition.

3 Methods

The automatic and intelligent computational algorithm was proposed which can be used for domains prediction based on the primary structure of protein. The DomAn approach predicts protein domains using a combination of information in the form of patterns, fragments and segments. Fragments (often called also as linkers) are a small regions of the protein chain that are outside of the domain. Segments are continuous sequence regions of domains and are used for uncontinuous domains representation. Patterns are a subsequence of the primary structure of the known protein which is created on each domain boundary of protein.

Proposed approach in the first step tries to match the longest patterns to specific protein sequence given as input. After that, the analysis of fragments occurs. In the last stage of the DomAn method, existence of discontinuous domains are checked using the database of segments.

In general the DomAn approach consists of two main elements:

- the proteins database contains generalized information - domain boundaries represented as patterns, fragments and segments templates - collected from different existing databases of protein domains boundaries (CATH [1, 2], DALI [3], PFAM [7], SCOP [4, 5, 6]) and PDB [8, 9, 10],
- a searching algorithm matched patterns, fragments, and segments to the unknown sequence and than based on decision scheme recognize where domains are situated

4 Implementation and Tests

The DomAn algorithm was tested in the CASP7 competition [11] and the DomAnS server participated in the domains boundaries definition category.

The CASP7 targets identifiers for which the DomAnS server obtained a result which was:

- the best and better than results of all others domains boundaries prediction servers (14): T0287, T0300, T0304, T0309, T0311, T0326, T0327, T0335, T0351, T0353, T0357, T0360, T0366, T0382.
- the best but could be equal to result of at least one of all others domains boundaries prediction servers (15): T0285, T0288, T0290, T0293, T0307, T0308, T0314, T0315, T0346, T0358, T0361, T0365, T0367, T0369, T0385. One can see that the DomAn approach achieved satisfactory solutions and is a very good

choice for those who want to predict domains boundaries in proteins based only on theirs primary structure.

5 Discussion

A new method of protein domains prediction has been proposed. Results proved usefulness of the proposed approach but it has to be improved in the future. Proteins database and automated mechanisms of the proteins database update should be proposed and developed. Also decision scheme and rules should be improved. In order to minimize the processing time of the DomAn algorithm the parallel version of this approach could be proposed.

References

[1] C.A. Orengo, A.D. Michie, S. Jones, D.T. Jones, M.B. Swindells, J.M. Thornton CATH - A Hierarchic Classification of Protein Domain Structures. *Structure*, 5,8:1093–1108, 1997.

[2] F.M.Pearl, C.F. Bennett, J.E. Bray, A.P. Harrison, N. Martin, A. Shepherd, I. Sillitoe, J. Thornton, C.A. Orengo, The CATH database: an extended protein family resource for structural and functional genomics. *Nucleic Acids Research*, 31,1:452–455, 2003.

[3] L. Holm, C. Sander. Touring protein fold space with Dali/FSSP. *Nucleic Acids Research*, 26:316–319, 1998.

[4] A.G. Murzin, S.E. Brenner, T. Hubbard, C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, 247:536–540, 1995.

[5] L. Lo Conte , S.E. Brenner , T.J.P. Hubbard , C. Chothia , A. Murzin SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Research.*, 30(1):264–267, 1995.

[6] A. Andreeva, D. Howorth , S.E. Brenner, T.J.P. Hubbard, C. Chothia, A.G. Murzin SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Research.*, 32:D226–D229, 2004.

[7] R.D. Finn , M. Marshall , A. Bateman iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions *Bioinformatics.*, 21:410–412, 2005.

[8] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissing, I.N. Shindyaloy, P.E. Bourne The Protein Data Bank. *Nucleic Acids Research.*, 28(1):235–242, 2000.

[9] J. Westbrook, Z. Feng, S. Jain, T.N. Bhat, N. Thanki, V. Ravichandran, G.L. Gilliland, W. Bluhm, H. Weissig, D.S. Greer, P.E. Bourne, H.M. Berman The Protein Data Bank: unifying the archive. *Nucleic Acids Research.*, 30(1):245–248, 2002.

[10] H.M. Berman, T. Battistuz, T.N. Bhat, W.F. Bluhm, P.E. Bourne, K. Burkhardt, Z. Feng, G.L. Gilliland, L. Iype, S. Jain, P. Fagan, J. Marvin, D. Padilla, V. Ravichandran, B. Schneider, N. Thanki, H. Weissig, J.D. Westbrook, C. Zardecki The Protein Data Bank. *Acta Cryst.*, D58:899–907, 2002.

[11] C.H. Tai, W.J. Lee, J. Vincent, B.K. Lee Assessment of Domain Predictions. *Assessors' Talks on Sixth Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction.*, Domain Boundaries (presented by Lee, B.K.)

[12] J. Blazewicz, P. Lukasiak, M. Milostan Some operations research methods for analyzing protein sequences and strutures. *4OR: Quarterly Journal of Operations Research.*, 4(2):91–123, 2006

Graph clustering approach for protein domain decomposition

***Maciej Milostan*[1,2], *Jacek Blazewicz*[2,3], *Piotr Lukasiak*[2,3],**

***Maciej Antczak*[2]**

1 Introduction

Proteins are one of the most important building blocks of life, they catalyze the necessary reactions in the cells to sustain life and improve metabolism. In many cases they used to function as exquisitely specific enzymes (catalysts). Determination of a native folded structure of a particular protein is a key to understand its function [1]. Such a determination, in most cases, can be done experimentally via crystallography [2] or NMR techniques [3]. If one knows the three dimensional structure then it is possible to identify its stable substructures called *domains*. Domains constitute functions of the protein. Automatic recognition of them makes function prediction easier. In the following sections novel algorithm for decomposing tertiary structure into domains has been proposed.

2 Problem formulation

Domains are semi-independent structures, so there exists interactions (e.g. chemical bonds) among amino acids that compose them. The basic idea is to recognize these interactions or spatial contacts and decompose protein on that basis.

The perfect algorithm should give an answer for a question: how to split the given three dimensional protein structure into domains. The input of the algorithm should be the tertiary structure and its output should be domains' assignments.

Although it is hard to define domain as a formal entity, it is possible to provide some basic features of the valid domain. A domain should have at least 40 residues, be

[1]E-mail: mmilostan@cs.put.poznan.pl
[2]Institute of Computing Science, Poznan University of Technology, Poznan, Poland
[3]Institute of Bio-organic Chemistry, Polish Academy of Sciences, Poznan, Poland

compact, have small cross-domain interface and not too many segments. Segment is a fragment of sequence composing part of a domain (c.f. [4] and [5]).

3 Method

The most straightforward approach is to represent protein structure as a graph of contacts. In such a case one has to identify contacts and then convert each residue in protein chain into vertex in the graph, and represent each contact as an edge. For purpose of contact identification distances between geometrical centers of sidechains and distances between C_α carbons have been used.

Given the protein graph one can apply graph clustering approaches for determination of potential domains.

Most efficient clustering methods need prior knowledge about number of clusters. To overcome this problem the idea of the structure coloring using simple rules has been proposed.

The proposed method contains following steps: contact graph generation, identification of small stable substructures, merging these substructures into clusters and final refinement of the assignments.

4 Experiments and Results

Exemplary results showed that proposed method has large potential. For the test set presented in [5] and [6] the algorithm gives similar results to one of the other compared approaches or SCOP [7, 8] database. This database contains information about domains boundaries and references to corresponding structures in PDB.

5 Conclusion

The method gives comparable results with other approaches known from literature but has lower complexity. Thus, further improvements are worth to be considered.

References

[1] J. Blazewicz and P. Lukasiak and M. Milostan Some operations research methods for analyzing protein sequences and structures. *4OR: A Quarterly Journal of Operations Research*, 4,2:91–123, 2006.

[2] J. Drenth Principles of protein X-ray crystallography. *Springer-Verlag Inc. NY*, 1999.

[3] K. Wuthrich NMR of proteins and nucleic acids. *John Wiley & Sons*, 1986.

[4] L. Holm and C. Sander Parser for protein folding units. *Proteins: Structure, Function, and Genetics*, 19:256–268, 1994.

[5] Y. Xu and D. Xu and HN. Gabow Protein domain decomposition using a graph-theoretic approach. *Bioinformatics.*, 16,12:1091-Ű1104, 2000.

[6] S. Jones, M. Stewart, A. Michie, M.B. Swindels, C. Orengo and J.M. Thornton Domain assignments for protein structures using a consensus approach: characterization and analysis. *Protein Sci.*, 7:233–242, 1998.

[7] A.G. Murzin, S.E. Brenner, T. Hubbard and C. Chothia SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology.*, 247:536–540, 1995.

[8] A. Andreeva, D. Howorth, S.E. Brenner, T.J.P. Hubbard, C. Chothia and A.G. Murzin SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acid Research.*, 32:226–229, 2004.

TU Clausthal
ICOLE 2007, LESSACH, AUSTRIA

RNA FRABASE – a database of RNA fragments

Marek Blazewicz[1,2]***, Mariusz Popenda***[3]***, Ryszard W. Adamiak***[3]***,***

Marta Szachniuk[2,3]

1 Introduction

The ultimate goal of bioinformatics is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology could be discerned. Three important sub-disciplines can be distinguished within this field of science: the development of new algorithms and statistics for associating relationships among members of large data sets, the analysis and interpretation of various types of data including nucleotide and amino acid sequences, protein domains, and protein structures, and, finally, the development and implementation of tools that allow for an efficient access and management of different types of information (Lesk, 2002). Designing and creating databases to store large amounts of biological data of different types is one of the most crucial task in the third sub-discipline mentioned above. There is a selection of different databases concerning the field of structural biology: Protein Data Bank, Nucleid Acid Database, The RNA Structure Database, SCOR and many others. Each of them stores some specific information about structures of proteins, DNAs and RNAs. Here, we present a novel database, named RNA FRABASE, devoted to RNA molecules. It is especially useful in case of studying RNAs containing particular secondary structure motifs. The database is also planned to be a component of the system for RNA tertiary structure prediction (Popenda et al., 2006).

[1]E-mail: mblazewicz@cs.put.poznan.pl
[2]Institute of Computing Science, Poznan University of Technology, Poznan, Poland
[3]Institute of Bio-organic Chemistry, Polish Academy of Sciences, Poznan, Poland

2 Methods

Molecular structure of RNAs, as well as of other biomolecules, can be described on several different levels. Sequence of residues forming RNA strand, called primary structure of the molecule, is its most general description. Next level is a secondary structure which describes single and double-stranded fragments within a molecule. These fragments form secondary structure motifs - like duplexes, junctions, loops, bulges, etc. - of different significance and functionality within living organisms. Finally, three-dimensional shape of a molecule is called its tertiary structure. RNA FRABASE contains the information about RNA fragments coming from RNAs with already solved structures. In the database primary structure is defined by the sequence of letters over the four-letter alphabet A,C,G,U, secondary structure is represented by Vienna notation (Hofacker et al., 1994). Tertiary structure is defined by torsional angles, base-pairs and atom coordinates. Primary and secondary structures of RNAs have been backcalculated with basic Linux scripts (awk, grep, etc.) and RnaView software tool (Yang et al., 2003) from tertiary structure descriptions given in PDB files. User obtains the required information from the database using the Web application implemented in PHP5. Query can be specified either by primary structure defined by IUPAC codes and/or secondary structure given in Vienna notation. As the result one receives sequence, two dimensional structure, experimental method, detailed information about torsional angles, base-pairs, web link to PDB file with atom coordinates and other information. Structures obtained match exactly the ones which are queried. Database have been created in PostgreSQL.

3 Summary

RNA FRABASE is a database with the Web interface. It provides a possibility to make an exhaustive search for 3D RNA fragments defined by specified structural parameters entered by the user. Actually the database contains information about 650 RNA structures and it is periodically updated by automatic procedures. RNA FRABASE has been already used by the scientists doing their research in the field of structural biology.

References

[1] A. M. Lesk. Introduction to Bioinformatics. Oxford University Press, 2002.

[2] M. Popenda, L. Bielecki, R. W. Adamiak. High-throughput method for the prediction of low-resolution three-dimensional RNA structures. *Nucleic Acids Symposium Series*, 50: 67-68, 2006.

[3] IUPAC codes: http://www.chem.qmul.ac.uk/iupac

[4] I. L. Hofacker, W. Fontana, P. F. Stadler, S. Bonhoeffer, M. Tacker, P. Schuster Fast folding and comparison of RNA secondary structures. *Monatshefte Chemie*, 125: 167-188, 1994.

[5] H. Yang, F. Jossinet, N. Leontis, L. Chen, J. Westbrook, H. M. Berman, E. Westhof Tools for the automatic identification and classification of RNA base pairs. *Nucleic Acids Research*, 31(13): 3450-3460, 2003.

SiteSeeker - An advanced motif discovery tool

***Klaus Ecker*[1,2]*, Lonnie Welch*[2]*, Dazhang Gu*[2]

1 Introduction

The research addressed with the SiteSeeker[3] is focused on the detection of motifs in the promoter regions of genes associated with specific problems.

By providing a workbench to try out various concepts for motif scoring and discovery strategies it is possible to create a motif discovery tool that combines advantages of existing solutions and thus outperforms them. The algorithmic objective is identifying one or more blocks of aligned words with low Hamming distance, low log likelihood, or large Entropy, or various combinations thereof.

The SiteSeeker is validated through the detection of promoters involved in the translation of gravi-stimulation protein. The SiteSeeker provides an iterative program execution but is currently dependent on direct user control.

The implemented solution can search for patters and can thus be applied to the checking of known motifs, in case of *Arabidopsis thaliana* from the AGRIS database, and the discovery of new motif candidates for a block of functionally related sequences.

2 Methods

The tool proposed here is intended to solve the problem of discovering a common motif in a given set of sequences. Hereby it is assumed that each sequence has the motif. A problem variation not considered here would be that only some of the sequences possess a common motif, in which case not only the motif has to be discovered, but also all the related sequences have to de identified. We allow an unbounded number of sequences. The strategy is to select a small number of sequences (up to 25), search these for a block of words considered as a motif candidate, and check if the motif is in all the other sequences.

[1]E-mail: ecker@ohio.edu

[2]Center for Intelligent, Distributed and Dependable Systems, Ohio University, Athens, USA

[3]The source code, executables and documents can be downloaded from http:ohio.educiddsmotif_discovery

The primary purpose of SiteSeeker is to create a workbench to try out various concepts for discovery strategies and motif scoring. The intention was to develop a tool that enables testing of different concepts regarding combined scoring methods and search strategies, that eventually ends in tool of practical use.

As was pointed out in a number of investigations, combining different scoring methods can lead to better results [HLK05, HJ06, LT06, TLB+05]. This is justified by the observation that each scoring method has its pros and cons. For example, it has been observed that Gibbs sampling produces in general too many false positives. The reason is that Gibbs sampler easily gets trapped in a local optimum (which is very likely a false positive). On the other hand, string comparison based on Hamming distance is often too restrictive.

Therefore, the central idea we are following in our algorithmic approach is identifying one or more blocks of aligned words, scored by the *bi-criterion* of simultaneously minimizing Hamming-2 distance (which is the sum of Hamming distances of promoter pairs) and maximizing entropy. This Pareto optimization problem returns a set of optimal or near optimal solutions. It is expected that this concept, enhanced by a few other improvements explained below will lead to a motif discovery program that outperforms the known tools.

The proposed solution uses a combination of exhaustive search and Gibbs sampling. First, two arbitrarily chosen sequences of the given set are scanned exhaustively to present the 100 "best" pairs of words, in the sense of one of the evaluation functions: minimum Hamming-2 distance, minimum likelihood, or maximum entropy. This takes little time (in the order of seconds) because there are only two sequences considered. Then for each solution pair, each of the other sequences is scanned for a word best matching the pair. By this we obtain for each of the other sequences a word located in some position, and all these positions together define an alignment for the sequences. Thus 100 blocks (each is specified by the start addresses of the block words) are generated in this step. So far the algorithm works deterministically.

For calculating the Hamming-2 distance of a pair of sequences, the scoring of nucleotide pairs is done on the basis of a 4 x 4-matrix Ω. Originally the elements of Ω would be 0 in the main diagonal and 1 elsewhere, reflecting the assumption that only mismatches add 1 to the Hamming distance. As is well-known, base pairs can have very different background frequencies (A and T usually have a much higher frequency than C and G). Since the motif search follows the strategy of identifying unexpected alignments, we could, modifying an idea of Zaslavsky et al. [ZS06], penalize pairs of matching symbols by replacing the zeroes in the main diagonal by values proportional to the background frequencies of the base pairs. The matrix Ω can be chosen individually and changed at any time, in order to fine-tune the resulting motif candidates.

Though we can choose from three evaluation functions in the exhaustive search phase, from experiments we could see that Hamming-2 distance mostly gives the best results. Comparison with known motifs showed that this part of the algorithm already identifies real motifs with remarkable success rate.

The Gibbs sampler can either start with one of the computed alignments of the Ham-

ming minimization step to search the neighborhood (defined by a slight modification of the alignment addresses), or it can start from randomly chosen alignments. For each block the neighborhood search is repeated up to 100 times, each time with a new seed alignment from the neighborhood, in the hope to escape local optima. The Gibbs sampler is able to improve the motifs from the Hamming minimization, add new solutions to the previously found, and may even discard solutions from the found set. Since our intention is of identifying the most unexpected alignment blocks in the given sequences, we pursue two objectives during this search: minimizing Hamming-2 distance and maximizing entropy. This offers a way to speed up the Gibbs sampling phase, because each block being worse in both, the Hamming-2 distance and the entropy than another block can be discarded from further consideration. As soon as no additional solutions can be found the algorithm terminates and provides a list of blocks regarded as the best (notice that they are still sub-optimal solutions).

Especially for the first program part, there are two thresholds that allow creating a small number of motif candidates of high quality. The first is the maximum allowed Hamming-2 distance of the two selected sequences, and the second restricts the maximum allowed Hamming-2 distance in the alignment of all sequences.

The tool does not provide any hints for the best motif length. Rather it is up to the user to run the program with different lengths and use his expertise to make the most reasonable choices.

A number of parameters can be set to control the program functions: The motif length can be changed any time. A switch is available for using or not using the matrix Ω for scoring pairs of nucleotides in two aligned sequences. Moreover the matrix Ω can be changed any time. It should be pointed out that the concept of this low-level scoring of aligned pairs of nucleotides turns out to be very useful as it takes care of the background distribution, and allows - at least up to some extend - to model nucleotide binding energies [WS00].

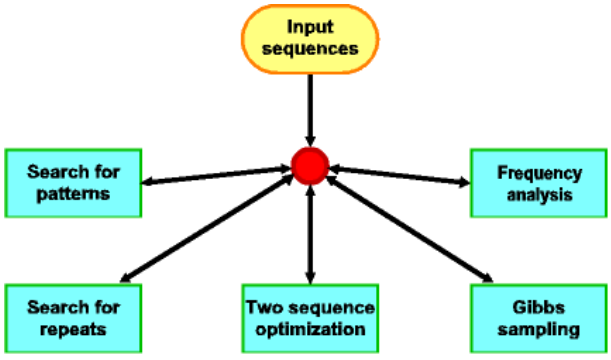

Figure 3: Architecture of SiteSeeker

The architecture of the SiteSeeker is depicted in Figure 3. It shows five components of which the **Two sequence optimization** and **Gibbs sampling** are the most important ones. Though there is no restriction in the order of execution, we recommend to first use **Two sequence optimization** to find quickly a promising set of motif candidates. The solutions may then be improved by the **Gibbs sampling** component. From the other modules, **Search for patterns** allows easy checking all the sequences for the putative motif. The program can handle up to 25 sequences. If the input has more sequences, the set should be divided in two or more parts. Once a number of motif candidates are found for one subset, the function **Search for repeats** checks the motif candidates on the other sequences. As to the other functions, **Frequency analysis** serves statistical questions, and **Search for repeats** finds all repeated words of a specified minimum length in a given sequence. Each of these steps has to be considered carefully and decisions for the next step depend on the current results. Currently we are developing a program version that automates the motif discovery process.

3 Results

After having checked for known motifs, we check if the sequences have a common exactly matching word. Of course, if there is one, then any two of the sequences will have the same word. So we choose two sequences and search for a pair of longest words of Hamming distance 0. If the length is sufficiently long (a common choice could be $>$ 8) one would then check if the other sequences contain the same word. This is already a good indication of a common motif, because it is very unlikely that even only two sequences of length = 3000 bp share a motif of length $>$ 8.

To validate SiteSeeker a group of 95 promoters from Arabidopsis Thaliana are chosen. Each group has between 1 and $\approx$100 promoters, and it is known from experiments that all promoters of a group share a common motif. We checked the first 40 groups. Four groups could not be solved because they contained only one promoter. In such case, of course, the tool fails because it needs at least two promoter sequences for comparison. Of the other 36 groups we were able to identify the motifs in 34 of them.

If not all of the other sequences have the same word there are these possibilities:

- the common word is too short (length $\leq$ 8) to allow for a trustworthy decision.
- the motif is not exactly matching; its structure is described by a regular expression of by a profile matrix,
- only part of the promoter set belongs to co-regulated genes.
- matches are coincidences with no visible practical meaning.

These questions should make clear why we consider the tool still as preliminary. When running the tool, too many choices and decisions have to be made at different stages to guide the program to significant results.

References

[HJ06] L.S. Hon, A.H. Jain. A deterministic motif finding algorithm with application to the human genome. *Bioinformatics.*, 22:1047–1054, 2006.

[HLK05] J. Hu, B. Li, D. Kihara. Limitations and potentials of current motif discovery algorithms. *Nucleic Acids Research.*, 33:4899–4913, 2005.

[LT06] N. Li, M. Tompa. Analysis of computational approaches for motif discovery. *Algorithms for Molecular Biology.*, I:8, 2006.

[TLB+05] M. Tompa, N. Li, T.L. Bailey, et al. Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biology.*, 23:137–144, 2005.

[WS00] C. Workman, G. Stormo ANN-SPEC: A Method For Discovering Transcription Factor Binding Sites With Improved Specificity. *Pac. Symp. Biocomput.*, 467–478, 2000.

[ZS06] E. Zaslavsky, M. Singh A combinatorial optimization approach for diverse motif finding applications *Algorithms for Molecular Biology.*, 1, 2006.

TU Clausthal
ICOLE 2007, LESSACH, AUSTRIA

Recent developments of LVQ

Barbara Hammer[1,2]

1 Machine learning and learning vector quantization

Machine learning (ML) deals with the task to improve the behavior of a machine based on experience. This principle can be applied in unsupervised scenarios or data mining where given data has to be processed without any additional knowledge about the relevant information to be extracted: prototypical techniques include the self-organizing map or neural gas. Alternatively, machine learning is applied in supervised scenarios where a regression task or classification has to be learned, i.e. data and the corresponding class labels are available, as for learning vector quantization (LVQ). A variety of different ML techniques exist including logical ones such as decision trees or statistical pattern recognition such as the support vector machine (SVM), which is one of the most powerful ML techniques available today.

One drawback of statistical machine learners consists in the fact that the methods are often black-box mechanisms which cannot easily be interpreted by humans. Prototype-based methods like LVQ constitute one exception: assume input data $\vec{x}_i$ are contained in $\mathbb{R}^n$. Then a prototype-based classifier is represented by a number of prototypes $\vec{w}_j \in \mathbb{R}^n$, i.e. the representation of the classifier is located in the same space as the data and can directly be inspected. Any prototype represents its receptive field, i.e. the data points closest to the prototype. A classification is determined by class labels Y_i attached to the prototypes, the classification function is $\vec{x} \mapsto Y_j$ where $\|\vec{x} - \vec{w}_j\|^2$ is minimum.

LVQ also provides a technique to learn the prototype locations based on given training data. Given a data point, LVQ 1 iteratively adapts the respective closest prototype toward (away from) the data point depending on whether the classification is correct (wrong). Similarly, LVQ 2.1 always adapts the closest correct and the closest wrong prototype given a data point. All classical variants of LVQ, however, are based on heuristics only and neither convergence nor the generalization ability are fully investigated for these basic methods. In recent years, a variety of extensions of LVQ have been proposed which, on the one side, provide a well-founded theoretical background

[1]E-mail: hammer@in.tu-clausthal.de
[2]Department of Informatics, Clausthal University of Technology, Clausthal-Zellerfeld, Germany

for LVQ, on the other side, extend the heuristics to powerful state-of-the-art methods which can compete with ML techniques such as SVM, while keeping the simplicity and interpretability of original LVQ.

2 Mathematical foundation

In recent years, the exact behavior of LVQ-type learning algorithms has been investigated by means of the theory of online-learning, see e.g. [1, 2, 3]. The theory of online learning allows to exactly characterize the typical dynamics of learning rules of specific type with respect to their mean behavior in typical model situations using methods from statistical physics. For this purpose, characteristic quantities of the system are identified and their dynamics is described in a learning setting with i.i.d. data in the limit of infinite data dimensionality. In this limit, the behavior can be exactly characterized by standard differential equations (ODE) under certain conditions, i.e. the evolution of characteristic quantities such as the size of the prototypes and the generalization error can be obtained by analytical or numeric integration of coupled ODEs. In [1, 2, 3], this procedure allows insight into typical behavior of methods such as LVQ 1, LVQ 2.1, and related methods: LVQ 1 shows surprisingly good generalization ability whereas LVQ 2.1 (even with window rule) diverges in almost all realistic scenarios.

An alternative method to provide a formal base is offered by the development of cost functions which are optimized by means of a stochastic gradient descent. This procedure yields LVQ-type learning rules. One particularly effective approach is built on the objective to obtain optimum generalization ability. The generalization ability of LVQ can be upper bounded for every possible learning scenario using PAC-style arguments based on the Rademacher complexity of LVQ-function classes. As shown in [4], the worst case generalization ability can be limited by the empirical error and a bound which includes the co-called hypothesis margin of LVQ-networks. This bound gives rise to a cost function: generalized LVQ (GLVQ) optimizes the ratio of the hypothesis margin and the sum of distances to the closest correct and wrong prototype [5, 6]. Interestingly, the generalization bounds show a strong regularization of LVQ-networks: the generalization error depends on the margin (which is optimized by GLVQ) and not the input dimensionality, i.e. the number of free parameters of LVQ-networks. This fact makes LVQ-networks ideally suited for situations where only few training data and high input dimensionality is present, provided the representation capability of LVQ networks is powerful enough to capture the underlying data regularity.

3 Adaptive metric

LVQ heavily relies on the euclidean metric. This fact causes severe problems for high dimensional or noisy data: errors likely accumulate such that the result of a euclidean comparison is almost random. Hence the representation capability of an LVQ network which is based on the euclidean metric is rather limited. The derivation of GLVQ by

means of a cost function provides a simple solution to this problem: one can substitute the euclidean metric by any differentiable similarity measure suited for the problem at hand [5]. Thereby, metric parameters can be kept adaptive and learned by means of a stochastic gradient descent together with the prototypes such that optimum classification and generalization ability is achieved. A very simple but powerful metric is the scaled euclidean metric which includes a relevance term per input dimension, i.e. a scaling factor which weights the relevance of every input dimension [6]. This extension has the benefit that the computational effort is only slightly increased and the excellent generalization ability of standard GLVQ is preserved [4]. Further, the relevance profile can directly be interpreted by humans. This extension, GRLVQ, has been successfully applied in a variety of applications including remote sensing image analysis [7] and clinical proteomics [8, 9, 10]. Extensions to even more general metrics can further improve the results depending on the area of application: spatial or temporal dependencies can be accounted for by adaptive local correlation measures as demonstrated in [5, 11] for various problems in bioinformatics and time series prediction. A correlation measure is beneficial to catch the overall development of data e.g. in macroarray analysis [12], and the incorporation of the full correlation of data can be included using a full matrix, as recently demonstrated in [13].

References

[1] M. Biehl, A. Ghosh, and B. Hammer. Dynamics and generalization ability of LVQ algorithms. *Journal of Machine Learning Research*, 8:323–360, 2007.

[2] A. Ghosh, M. Biehl, and B. Hammer. Performance analysis of LVQ algorithms: a statistical physics approach. *Neural Networks*, 19:817–829, 2006.

[3] M. Biehl, A. Ghosh, and B. Hammer. Learning vector quantization: The dynamics of winner-takes-all algorithms. *Neurocomputing*, 69(7-9):660–670, 2006.

[4] B. Hammer, M. Strickert, and T. Villmann. On the generalization ability of GRLVQ networks. *Neural Processing Letters*, 21(2):109–120, 2005.

[5] B. Hammer, M. Strickert, and T. Villmann. Supervised neural gas with general similarity measure. *Neural Processing Letters*, 21(1):21–44, 2005.

[6] B. Hammer and Th. Villmann. Generalized relevance learning vector quantization. *Neural Networks*, 15(8-9):1059–1068, 2002.

[7] T. Villmann, E. Merényi, and B. Hammer. Neural maps in remote sensing image analysis. *Neural Networks*, 16(3-4):389–403, 2003.

[8] T. Villmann, F.-M. Schleif, and B. Hammer. Prototype-based fuzzy classification with local relevance for proteomics. *Neurocomputing*, 69:2425–2428, 2006.

[9] F.-M. Schleif, B. Hammer, and Th. Villmann. Margin based active learning for LVQ networks. *Neurocomputing*, 70(7-9):1215–1224, 2007.

[10] F.-M. Schleif, Th. Villmann, and B. Hammer. Prototype based fuzzy classification in clinical proteomics. *International Journal of Approximate Reasoning*, to appear.

[11] B. Hammer, M. Strickert, and T. Villmann. Prototype based recognition of splice sites. In U. Seiffert, L.C. Jain, and P. Schweitzer, editors, *Bioinformatics using computational intelligence paradigms*, pages 25–55. Springer, 2005.

[12] M. Strickert, U. Seiffert, N. Sreenivasulu, W. Weschke, T. Villmann, and B. Hammer. Generalized relevance LVQ (GRLVQ) with correlation measures for gene expression analysis. *Neurocomputing*, 69(6–7):651–659, March 2006. ISSN: 0925-2312.

[13] P. Schneider, M. Biehl, and B. Hammer. Relevance matrices in LVQ. In M. Verleysen, editor, *Proc. Of European Symposium on Artificial Neural Networks (ESANN'2006)*, pages 37–42, Brussels, Belgium, 2007. d-side publications.

TU Clausthal
ICOLE 2007, LESSACH, AUSTRIA

Competitive learning with conscience

Sebastian Magnus[1,2]

1 Introduction

Vector quantization (VQ) is used for classification, pattern identification, clustering or compression in many fields. VQ distributes prototypes at representative positions of the data space. In applications such as data compression, the density of neurons should meet the density of input data. So every neuron should win the competition for an input with the same probability. However, popular VQ schemes such as k-means do not find such an allocation. Here, contrasts are dampened, such that rare examples are learned with a lower priority and neurons move to regions with higher density.

Model	**Magnification**
K-Means	$d/(d+2)$
NG	$d/(d+2)$
SOM	$(1+12M2(\sigma))\,/\,(3+18M2(\sigma))$
SOM: $1 << \sigma << N$	$2/3$
SOM: small σ	$1/3$
d: intrinsic dimension σ: Neigborhood range	

2 Magnification and Magnification Control

Magnification describes the relation of the input density $P(w)$ and the neuron density $\rho(w)$. This relation is usually expressed by a power law $P(w)\sim\rho(w)^a$. The included table shows the magnification factor for common algorithms. Here, the factor is smaller than one. Contrasts are dampened and rare examples have a lower influence. If we had a magnification factor greater than one, contrasts would be emphasized and rare

[1]E-mail: sebastian.magnus@tu-clausthal.de
[2]Department of Informatics, Clausthal University of Technology, Clausthal-Zellerfeld, Germany

samples would have a higher effect. This can sometimes be useful, for example to detect abnormal system behaviour, rare ground patterns or to explore unknown data spaces. A magnification factor of 1 corresponds to a perfect match of prototype allocation and data distribution.

Exact approaches to investigate and control the magnification exist, but they have crucial disadvantages. In local learning, the local density is estimated. This estimation is very costly and may lead to numerical problems. Convex/concave learning modifies the distance by a power law. This gives explicit magnification control but is very unstable. Winner relaxing learning assigns a different learning rate to the winner neuron and is also very unstable.

Conscience has been proposed by deSieno to achieve a magnification factor 1 for online self-organizing maps. Frequent winners get a penalty whereas rare winners are boosted. This apporach has not been applied to alternative VQ schemes such as k-means or batch SOM. In this contribution, we show the effectiveness of conscience learning for (batch) k-means clustering.

3 The classical K-Means

K-Means attaches inputs to their nearest neuron. Then the weights of the neurons are set to the center of gravity of their attached inputs. The learning loop is repeated until there are no further changes in the allocation.

Initialize
Repeat:

$$d_{ij} = |w_i - x_j|$$

$$y_{ij} = \begin{cases} 1 & \text{if } d_{ij} < d_{kj} \forall k \neq i \\ 0 & \text{otherwise} \end{cases}$$

$$w_i = \sum_j \frac{y_{ij} \cdot x_j}{\sum_j y_{ij}}$$

K-Means is very fast, easy to implement and it always converges. The main disadvantage is that it does not consider the input density and it depends on the initialization. The neurons are rather equidistant than equiprobably distributed. So K-Means is very inefficient in respect to data compression.

4 Conscience and Conscience SOM by DeSieno

DeSieno proposed to use a conscience term for selforganizing feature maps to achieve an equidistributed allocation of neuron weights. The experiments were restricted to one dimensional SOMs without neighborhood. First the distances without conscience are determined as usual. Every neuron gets a conscious term depending on how often it has won in competitions:

$$p_{i_{new}} = p_{i_{old}} + B(y_{ij} - p_{i_{old}}) \text{ with } 0 < B << 1$$

The conscience increases for the winner and decreases for all other neurons. With this conscience a bonus or penalty is determined:

$$b_i = C \cdot \left(\frac{1}{N} - p_i \right)$$

This term is substracted directly from the distance. A second competition is started to determine which neuron will be winner for adaption, this time using the conscience. Frequent loosers get a big bonus substracted from their distance, frequent winners get a smaller bonus or even a penalty. DeSieno reports that the use of conscience speeds up the training process and neurons can jump over gaps in input space. After training, every neuron wins about the same number of training inputs and has about the same probability of winning.

5 Conscience K-Means

The goal of conscience K-Means is to obtain both, the advantages of a fast and easy K-Means and the equiprobable neuron density which we achieve with exact magnification control only at high costs. The original K-Means is only slightly modified by adding a conscience term to the distance function. This term is determined by multiplying a constant conscience factor to the number of competitions already won by a neuron and dividing it by the number of total inputs.

Initialize
Repeat:

$$d_{ij} = |w_i - x_j| + C \frac{\text{inputs}_{\text{won}}}{\text{inputs}_{\text{total}}}$$

$$y_{ij} = \begin{cases} 1 & \text{if } d_{ij} < d_{kj} \forall k \neq i \\ 0 & \text{otherwise} \end{cases}$$

$$w_i = \sum_j \frac{y_{ij} \cdot x_j}{\sum_j y_{ij}}$$

First experiments show that the algorithm constitutes a very robust and efficient method to achieve magnification factor approximately one. The results of conscience K-Means can be further improved by taking the history of learning into account.

References

[1] D. DeSieno. Adding a conscience to competitive learning. *Proc. ICNN'88*, 117-124, 1988.

[2] T. Villmann, J. C. Claussen. Investigation of Magnification Control in Self-Organizing Maps and Neural Gas. *Neural Computation*, 18(2): 446-469, 2006.

[3] B. Hammer, A. Hasenfuss, T. Villmann. Magnification control for batch neural gas. *Neurocomputing*, 70(7-9): 1225-1234, 2007.

TU Clausthal
ICOLE 2007, LESSACH, AUSTRIA

SOM for biological uses – Motif discovery on genomic sequence of S. cerevisiae

Mirco Schindler[1,2,3]

1 Introduction

This contribution reviews an application of the Self-Organizing Map for biological uses based on the article of Mahony et al. First the Standard Self-Organizing Map (SOM) model and some general applications of it will be described. Then, a motif discovery method called SOMBRERO which is based on SOM will be presented. SOMBRERO is available free of charge from http://bioinf.nuigalway.ie/sombrero. The advantage of this approach is that it can be used to simultaneously characterize every feature present in the input data space.

2 Standard self-organizing map

The SOM has been introduced by Teuvo Kohonen in 1984. It is self-organized; this means that there is no external teacher and the network has to extract the relevant information from the data by itself. Assume a n-dimensional euclidean data space is given. The aim of SOM is to get a clustering of this input space. The result after the training phase of the SOM is that the neurons from the network are specialized on different clusters of the input space.

One neuron reacts on an input. This neuron which shows the maximum reaction is the so called winner-neuron. Further, the network is equipped with a regular, often two-dimensional mesh. The input data has n dimension, where n is often very large.

During training, this neighbourhood structure is used to achieve a topology representing mapping of the input space. In the standard SOM model each neuron in the network is characterized by a n dimensional weight vector. The dimension of the weight

[1]E-mail: mirco.schindler@tu-clausthal.de
[2]Department of Informatics, Clausthal University of Technology, Clausthal-Zellerfeld, Germany
[3]Please note that this abstract does not contain original work but it summarizes a method, SOMBRERO, developed by Mahony et al. as indicated in the refence.

vector is equivalent to the dimension of the input data. During the training the vector is modified according to the given data. The neuron grid is trained such that it will be spread out over the input space in a topology preserving way.

3 SOM training

For training, the winner neuron and its neighbourhood are adapted. There are many different definitions of neighbourhood adaptation which are mostly equivalent for training. Often, a Gaussian curve is used:

$$\varphi(n_j, n_{\text{winner}}) = e^{-\frac{(D(n_j, n_{\text{winner}}))^2}{\sigma}}$$

Here, the function D returns the distance between the winner-neuron n_{winner} and the neuron n_j. During the training phase, each weight vector evolves to represent a different region of the input data. Nodes that are located close to each other on the network will become similar weights. Before training, weight vectors have to be initialized. Often they are distributed randomly in the domain of the input data space. Afterwards, the learning rate η has to be set. During training a random input data from the input space is chosen. The distances of the weight vectors to the data are computed and the winner is determined. Then the weight vectors for all neurons are adapted according to the following formula, where ς denotes the input

$$w_i \leftarrow w_i + \eta\varphi(n_i, n_{\text{winner}}))(\xi - w_i)$$

After adapting the weight vectors the process is repeated and the variables σ and η are decreased. An alternative fast training method is offered by batch adaptation.

Applications of SOM include various data mining tasks. In 1996 WEBSOM was developed. In the WEBSOM method the SOM algorithm is used to automatically organize very large and high-dimensional collections of text documents. The result is a map which visualizes the context of the documents. Other applications include robotics, recognition of speech, and motif discovery on genomic sequences, which will be explained in more detail, now.

4 SOMBRERO

SOMBRERO (Self-Organizing Map for biological regulatory element recognition and ordering) is an implementation which is capable of discovering multiple distinct motifs present in a single data set. In the article from Mahony, SOMBRERO has been trained on genomic sequences from Saccharomyces cerevisiae. The input data is given by eight genomic sequences from S. cerevisiae. This is presented to SOM as l-mers with window size ranging from 16 to 22bp. After training, all neurons represent a characterization of the various motif features present in the input sequences.

5 Adaptation of the standard model

The distance function for SOM training has to be adapted according to the given task. A score function S(x) is used to rate the similarity of a string x and a motif defined by a neuron. This similarity is given by the log-likelihood ratio of a DNA string. Neurons are characterized by a profile weight matrix (PWM). They represent probabilities of bases. An entry of a PWM of node z looks like this:

$$f_{ib}^{z} = \frac{c_{ib}^{z}}{n_z}$$

with c_{ib}^{z} number of occurrences of base b at position i and $n_z = \sum_b c_{ib}^{z}$ at position i.

The initialization of the PWM's is not randomly. Each neuron represents a unique value: nodes have a preference for a certain base, as determined by the quadrant of the SOM. To train the SOM, the batch-version of the SOM algorithm is used. The neighbourhood function is given by the Gaussian neighbourhood. The training ends after 100 steps. After this the two-dimensional grid of PWM's represents a characterisation of the various motif features present in the input sequences. With other words each node will contain a different potential motif.

6 Motif identification and results

But how to find which neuron represents an interesting sequence? The focus of this approach is to find overrepresented motifs. For this purpose a third-order Markov model is used as background model. Using this background model each node z can be ranked with the score-function

$$z_{\text{score}} = \frac{n_z - < n_z >}{\sigma_z}$$

with standard deviation σ_z and the expected number of occurrences $< n_z >$.S

Three different methods (SOMBRERO, MEME, AlignACE) are compared in the article for motif discovery. Data are given by ten yeast promoter sequence datasets. The results of this study show that SOMBRERO yields less false negatives than the other methods in nine of the ten cases.

References

[1] Shaun Mahony, David Hendrix, Aaron Golden, Terry J. Simth, Daniel S. Rokhsar Transciption factor binding site identification using the Self-Organizing Map *Bioinformatics*, 21(9): 1807-1814, 2005.

[2] Promoter Database of S.cerevisiae (SCPD; http://cgsigma.cshl.org/jian/)

[3] Hughes J.D., Estep P.W. Tavazoie S. and Church G.M. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyes cerevisiae *J. Mol. Bio.*, 296, 1205-1214, 2000.

[4] Kohonen T. (1995) Self-Organizing Maps *Springer-Verlag, Berlin.*

TU Clausthal
ICOLE 2007, LESSACH, AUSTRIA

Visualising gene expression data

Bassam Mokbel[1,2,3]

1 Introduction

This lecture was about useful methods to cluster and visualise gene expression data. Two different approaches for dimension reduction were described and compared. First an adaptation of Multidimensional Scaling for high throughput and large data sets by Strickert et al. [1] has been presented. Secondly, the possibility to use Hyperbolic Self-Organizing Maps was discussed, referring to Ritter et al. [2] and Hammer et al. [3]. All methods can be used to cluster mostly unknown data, such as gene expression levels from series of DNA-micro- or macroarray experiments.

2 High-Throughput Multidimensional Scaling

Multidimensional Scaling (MDS) is a nonlinear technique for embedding high-dimensional data into a low-dimensional target space, most commonly the 2- or 3-dimensional Euclidean space for visualisation. For every input point, a low-dimensional data surrogate is placed in such a way that their mutual distances match the original ones in the input space as good as possible. Therefore their dissimilarity is measured with a stress function, which needs to be optimised with respect to good surrogate locations. This optimisation is realised by a stochastic gradient descent.

An adaptation of the classic MDS algorithm has been presented in [1], which is suited for very large data sets and is referred to as High-Throughput MDS (HiT-MDS). The authors achieved substantial speedup by using an efficient stress function with superior convergence properties based on the Pearson-correlation. Also, instead of recalculating the distance matrix correlation for each update step, it is locally incrementally adjusted. The new values are created from existing old ones additionally adding the n-1 specific changes caused by the adaptation of the current target point. Strickert et al.

[1]E-mail: bassam.mokbel@tu-clausthal.de
[2]Department of Informatics, Clausthal University of Technology, Clausthal-Zellerfeld, Germany
[3]Please note that this abstract does not present original work but it summarizes and discusses methods for data visualization developed by Strickert et al. and Ritter et al. as indicated in the references.

tested their method on sample data to cluster gene expression experiments, aiming at the visualisation of inter-experiment relationships. Overall results showed a higher confidence in reconstruction quality of the HiT-MDS embedding than the one achieved with Principal Component Analysis (PCA), a technique also common in this area. Furthermore HiT-MDS was used for coexpression analysis of DNA-macroarray data, tracking gene expression levels throughout several developmental stages, again with better results than with PCA.

3 Hyperbolic Self-Organising Map

The Hyperbolic Self-Organising Map (HSOM) is an adaptation of Kohonen's SOM introduced in [2], taking advantage of the features of the hyperbolic geometry and the Poincaré Disc Model for visualisation. The hyperbolic plane is characterised by a constant negative Gaussian curvature, which, modelled in 2-dimensional Euclidean space, offers great opportunities for visualising large maps. As it resembles the view through a fisheye camera lens, the Poincaré Disc shows nearly Euclidean behaviour in its center. Therefore, images in the focus of the disc appear rather normal, while everything gets condensed exponentially towards the circular border.

In the classic SOM, a lattice in the Euclidean space provides coordinates by which the neurons' positions are defined. Subsequently the lattice of the introduced HSOM must be a regular tessellation of the hyperbolic plane - in this case a triangulation. The only adaptation needed from the original SOM algorithm is therefore to calculate the neighbourhood of the node and the node distances according to the Riemannian metric of the hyperbolic grid.

The research of Ritter's group is accompanied by a project website[1] which includes a variety of interesting application examples, among them the clustering and visualisation of text and handwritten ciphers. In recent work presented in [3] the HSOM was further extended to the Relational HSOM to cluster gene expression data obtained at IPK Gatersleben[2]

4 Visualisation

Regarding visualisation, there is a natural trade-off between a simplified, more abstract model of the input data versus a more truthful representation with finer granularity. While an HSOM visualisation allows for intuitive, focused browsing even in very large maps and delivers an efficient hierarchical representation of information, a one-to-one correspondence between input data and their low-dimensional counterparts vanishes. The HiT-MDS method behaves contrarily, as its mapping depicts every single data point and thereby delivers more details about the data distribution and granularity. Although

[1] http://www.techfak.uni-bielefeld.de/ags/ni/projects/hsom/

[2] Institute of Plant Genetics and Crop Plant Research Gatersleben, http://www.ipk-gatersleben.de

browsing of large maps could be less intuitive due to the lack of a neighbourhood structure, the algorithm works fast for very large data sets, in opposition to an HSOM which is, as stated, relatively slow.

References

[1] M. Strickert, S. Teichmann, N. Sreenivasulu, U. Seifert. High-Throughput Multi-Dimensional Scaling (HiT-MDS) for cDNA-Array Expression Data. *Artificial Neural Networks: Biological Insprations ICANN 2005*, LNCS 3696: 625-634, 2005.

[2] H. Ritter. Self-Organizing Maps on non-euclidean Spaces. *Kohonen Maps*, 97-110, 1999.

[3] B. Hammer, A. Hasenfuss, F. Rossi, M. Strickert. Topographic Processing of Relational Data. *to appear at WSOM 2007*.

Time series clustering

Andrej Gisbrecht[1,2,3]

1 Introduction

The aim of this contribution is to summarize the results of two articles of the literature, the popular work 'Clustering of time series subsequences is meaningless' by Keogh and colleagues [2] and the article 'Clustering Short Time Series Gene Expression Data'by Ernst and colleagues. On the one hand, problems of moving-window clustering for long time series are explained, on the other hand, an efficient clustering approach for clustering short time series is discussed.

2 STS-Clustering is meaningless

Time series data is a frequent type of data which has to be clustered for data mining tasks, often. Clustering methods can be classified into two categories:

- Whole time series clustering: A set of time series is given, the objective is to group similar time series into clusters whereby time series are compared as a whole.
- Subsequence clustering: Given a single time series, subsequences (parts) are extracted with a sliding window and the single parts of the time series have to be clustered.

Both objectives can be frequently found in literature. Surprisingly, subsequence clustering likely yields meaningless results, although found in literature very often. Clustering time series as a whole has to take care for the specific shape of time series. Both issues will be discussed in the following, focussing first on subsequence clustering and addressing whole time series clustering later.

[1]E-mail: andrej.gisbrecht@tu-clausthal.de
[2]Department of Informatics, Clausthal University of Technology, Clausthal-Zellerfeld, Germany
[3]Please note that this abstract does not contain original work but summarizes results published by Keogh et al. and Ernst et al. as indicated in the references.

Definitions

At the beginning, a few fundamental definitions are made:

- Time series: A time series $T = t_1, ..., t_m$ is an ordered set of m real-valued variables.
- Subsequence: A subsequence of T is a sampling of length $w < m$ of contiguous positions of T.
- Sliding window: A matrix S of all possible subsequences can be built by 'sliding a window' along T and placing subsequence C_p in the p-th row of S

Meaninglessness of STS Clustering

The meaninglessness of subsequence clustering can be observed in experiments: simple clustering algorithms such as k-means are applied on stock market data and stored in a matrix X. This is performed several times with restarts. The same experiment is performed for random walk data, and the result is stored in a matrix Y. Let $D(X)$ be the average distance between each set of cluster centers in X. Let $D(X, Y)$ be the average distance between each set of cluster centers in X, to cluster centers in Y. Then cluster-meaninglessness can be defined as follows:

$$M(X, Y) = \frac{D(X)}{D(X, Y)}$$

To exclude that the result depends on specific values of the number of clusters k and stochastic clustering result, a set of different parameters should be used. Additionally, the result can be compared to clustering whole time series. In the results presented in [2], it can be seen that the cluster centers found by streaming time series clustering on any particular run of k-means on stock market dataset are not significantly more similar to each other than they are to cluster centers taken from random walk data.

A hidden constraint

This observation is quite surprising, and can be explained as follows, referred to by Keogh et al. as '*a hidden constraint*'.: If one uses k-means with $k = 1$ on any time series without trend, one gets a horizontal line as cluster center. Why does this happen? In the article [2], the following explanation is given: *Imagine an arbitrary datapoint t_i somewhere in the time series T which is away from the borders. If the time series is much longer than the window size, then virtually all datapoints are of this type. As the sliding window passes by, the datapoint first appears as the rightmost value in the window, then it goes on to appear exactly once in every possible location within the sliding window. So the datapoint contribution to the overall shape is the same on every position of the time window and a horizontal line results. The average of many*

horizontal lines is clearly just another horizontal line. For any dataset, the weighted (by cluster membership) average of k clusters sums up to the global mean. (see [2]) Thus, this hidden constraint limits the utility of streaming time series clustering.

3 Short time series gene clustering

Time series gene expression experiments constitute a popular method for studying biological processes. Since they require multiple microarrays, which are very expensive, most of them are short. Even if microarrays get cheaper in the future, short time series will remain important in studies where it is prohibitive to obtain large quantities of biological material (like blood probes). Microarray data have been investigated in the literature and some clustering algorithms exist which yield biological insights. However, these are not designed for time series data.

The paper [1] is based on basic profiles, that represent the behavior of genes. To find significant changes the raw expression values are converted into log ratios with respect to the first time point, so that the first value of the series is always 0. A parameter c is defined that controls the amount of change a gene can make between two successive time points. For n time points this strategy results in $(2c+1)^{n-1}$ distinct profiles. For six time points and $c = 2$ this method results in $5^5 = 3125$ model profiles. Each of those would cover only few genes, and the overall number of profiles would be too big to handle. Therefore, a manageable subset of the profiles is selected. This model profiles should represent all possible expression profiles as good as possible. So the problem occurs to select a set of m profiles such that the minimum distance between any two profiles is maximized. Unfortunately, this problem is NP-hard such that a greedy algorithm is proposed.

After choosing the model profiles, the gene expression profiles are assigned, such that the distances between gene profiles and model profiles becomes minimal. The number of genes assigned to model profile m_i is denoted as $t(m_i)$. To identify the significantly enriched profiles, one has to take into account, that many gene profiles rely on statistical effects. Therefore, a background model is considered: One assumes that there is no order between single experiments and the number of genes that each profile would get on average is counted. For n time points each gene has $n!$ possible permutations. For each possible permutation genes are assigned to their closest model profile. Let s_i^j be the number of genes assigned to model profile i in permutation j. Set $S_i = \sum_j s_i^j$. The expected number of genes for each profile model is $E_i = S_i/(n!)$. For significantly enriched model profiles m_i it holds $t(m_i) > S_i$.

The experiments presented in [1] with biological data shows that this algorithm finds small clusters even in large noisy datasets, i.e. reasonable clustering is possible in this situation.

References

[1] J. Ernst, G.J. Nau, Z. Bar-Joseph, Clustering Short Time Series Gene Expression Data, Bioinformatics 21(S1): i159-i168, 2005.

[2] E. Keogh, J. Lin, W. Truppel, Clustering of Time Series Subsequences is Meaningless: Implications for Previous and Future Research, ICDM 2003.

Relational Neural Gas

Alexander Hasenfuss[1,2]*, Barbara Hammer*[2]

1 Introduction

Topographic maps such as the self-organizing map (SOM) constitute a valuable tool for robust data inspection and data visualization which has been applied in diverse areas such as telecommunication, robotics, bioinformatics, business, etc. [7]. Alternative methods such as neural gas (NG) [10] provide an efficient clustering of data without fixing a prior lattice. This way, subsequent visualization such as multidimensional scaling [9] can readily be applied, whereby no prior restriction of a fixed lattice structure as for SOM is necessary and the risk of topographic errors is minimized. For NG, an optimum (nonregular) data topology is induced such that browsing in a neighborhood becomes directly possible [11].

In the last years, a variety of extensions of these methods has been proposed to deal with more general data structures. This accounts for the fact that more general metrics have to be used for complex data such as microarray data or DNA sequences. Further it might be the case that data are not embedded in a vector space at all, rather, pairwise similarities or dissimilarities are available.

In this contribution, two canonical approaches are presented to extend NG and other clustering methods to relational data given by pairwise similarities or dissimilarities, respectively. The method combines the well-founded mathematics of NG in terms of a cost function and its widespread and efficient applicability.

2 Neural gas

Neural clustering and topographic maps constitute effective methods for data preprocessing and visualization. Classical variants deal with vectorial data $\vec{x} \in \mathbb{R}^n$ which are distributed according to an underlying distribution P in the euclidean plane. The goal of neural clustering algorithms is to distribute prototypes $\vec{w}^i \in \mathbb{R}^n$, $i = 1, \ldots, k$

[1]E-mail: hasenfuss@in.tu-clausthal.de
[2]Department of Informatics, Clausthal University of Technology, Clausthal-Zellerfeld, Germany

among the data such that they represent the data as accurately as possible. A new data point $\vec{x}$ is assigned to the *winner* $\vec{w}^{I(\vec{x})}$ which is the prototype with smallest distance $\|\vec{w}^{I(\vec{x})} - \vec{x}\|^2$. This clusters the data space into the receptive fields of the prototypes.

The cost function of Neural Gas (NG) [10] is given by

$$E_{\mathrm{NG}}(\vec{w}) = \frac{1}{2C(\lambda)} \sum_{i=1}^{k} \int h_\lambda(k_i(\vec{x})) \cdot \|\vec{x} - \vec{w}^i\|^2 \, P(d\vec{x})$$

where

$$k_i(\vec{x}) = |\{\vec{w}^j \mid \|\vec{x} - \vec{w}^j\|^2 < \|\vec{x} - \vec{w}^i\|^2\}|$$

is the rank of the prototypes sorted according to the distances, $h_\lambda(t) = \exp(-t/\lambda)$ scales the neighborhood cooperation with neighborhood range $\lambda > 0$, and $C(\lambda)$ is the constant $\sum_{i=1}^{k} h_\lambda(k_i(\vec{x}))$. The neighborhood cooperation smoothes the data adaptation such that, on the one hand, sensitivity to initialization can be prevented, on the other hand, a data optimum topological ordering of prototypes is induced by linking the respective two best matching units for a given data point [11]. Classical NG is optimized in an online mode. For a fixed training set, an alternative fast batch optimization scheme can be derived which in turn computes ranks, which are treated as hidden variables of the cost function, and optimum prototype locations [1].

3 Relational data

Relational data x^i are not contained in a euclidean vector space, rather, pairwise similarities or dissimilarities are available. Batch optimization can be transferred to such situations using the so-called generalized median [1, 8]. Assume, distance information $d(x^i, x^j)$ is available for every pair of data points $x^1, \ldots, x^m$. Median clustering reduces prototype locations to data locations, i.e. adaptation of prototypes is not continuous but takes place within the space $\{x^1, \ldots, x^m\}$ given by the data. We write w^i to indicate that the prototypes need no longer be vectorial. For this restriction, the same cost function as beforehand can be defined whereby the euclidean distance $\|\vec{x}^j - \vec{w}^i\|^2$ is substituted by $d(x^j, w^i)^2 = d(x^j, x^{l_i})^2$ whereby $w^i = x^{l_i}$. Median clustering substitutes the assignment of $\vec{w}^i$ as (weighted) center of gravity of data points by an extensive search, setting w^i to the data points which optimize the respective cost function for fixed assignments. This procedure has been tested e.g. in [1, 2]. It has the drawback that prototypes have only few degrees of freedom if the training set is small. Thus, median clustering usually gives inferior results compared to the classical euclidean versions when applied in a euclidean setting.

Relational clustering constitutes a direct transfer of the standard euclidean training algorithm to more general settings allowing smooth updates of the solutions. The essential observation consists in a transformation of the cost functions as defined above to the so-called relational dual. Assume training data $x^1, \ldots, x^m$ are given in terms of

pairwise distances $d_{ij} = d(x^i, x^j)^2$. We assume that it originates from a Euclidean distance measure, that means, we are able to find (possibly high dimensional) Euclidean points $\vec{x}^i$ such that $d_{ij} = \|\vec{x}^i - \vec{x}^j\|^2$. Note that this notation includes a possibly non-linear mapping (feature map) $x^i \mapsto \vec{x}^i$ corresponding to the embedding in a Euclidean space. However, this embedding is not known, such that we cannot directly optimize the above cost functions in the embedding space. The key observation is based on the fact that optimum prototype locations $\vec{w}^j$ of NG can be expressed as linear combination of data points. Therefore, the unknown distances $\|x^j - w^i\|^2$ can be expressed in terms of known values d_{ij}: Since the prototypes can be expressed in terms of data points in the form of $\vec{w}^i = \sum_j \alpha_{ij}\vec{x}^j$ where $\sum_j \alpha_{ij} = 1$ the identity

$$\|\vec{w}^i - \vec{x}^j\|^2 = (D \cdot \alpha_i)_j - 1/2 \cdot \alpha_i^t \cdot D \cdot \alpha_i$$

holds where $D = (d_{ij})_{ij}$ constitutes the distance matrix and $\alpha_i = (\alpha_{ij})_j$ the coefficients. This yields an alternative algorithmic formulation where explicit prototypes are substituted by terms incorporating the variables α_{ij} and the known distance matric D.

The methods have successfully been aplied to different benchmark problems in protein classification, cytogenetics, and pattern recognition [4, 5]. Thereby, extensions such as supervision and magnification control [3, 6] can directly be integrated, because the principal theory of NG can directly be transferred to this setting.

References

[1] M. Cottrell, B. Hammer, A. Hasenfuss, and T. Villmann (2006), Batch and median neural gas, *Neural Networks*, 19:762-771.

[2] B. Hammer, A. Hasenfuss, F.-M. Schleif, and T. Villmann (2006), Supervised median neural gas, In Dagli, C., Buczak, A., Enke, D., Embrechts, A., and Ersoy, O. (Eds.), *Intelligent Engineering Systems Through Artificial Neural Networks* 16, Smart Engineering System Design, pp.623-633, ASME Press.

[3] B. Hammer, A. Hasenfuss, F.-M. Schleif, and T. Villmann (2006), Supervised batch neural gas, In *Proceedings of Conference Artificial Neural Networks in Pattern Recognition (ANNPR)*, F. Schwenker (ed.), Springer, pages 33-45.

[4] B. Hammer, A. Hasenfuss (2007), Relational Neural Gas, accepted for KI'07.

[5] A. Hasenfuss, B. Hammer (2007), Relational Topographic Maps, accepted for IDA'07.

[6] B. Hammer, A. Hasenfuss, and T. Villmann (2007), Magnification control for batch neural gas, *Neurocomputing* 70:1225-1234.

[7] T. Kohonen (1995), *Self-Organizing Maps*, Springer.

[8] T. Kohonen and P. Somervuo (2002), How to make large self-organizing maps for nonvectorial data, *Neural Networks* 15:945-952.

[9] J. B. Kruskal (1964), Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* 29:1-27.

[10] T. Martinetz, S.G. Berkovich, and K.J. Schulten (1993), 'Neural-gas' network for vector quantization and its application to time-series prediction, *IEEE Transactions on Neural Networks* 4:558-569.

[11] T. Martinetz and K. Schulten (1994), Topology representing networks. *Neural Networks* 7:507-522.

Multi-Agent-Systems challenges

Slawomir Deren[1,2]

1 Introduction

Multi-Agent-Systems (MAS) are systems composed of several interactive autonomous agents. In practice they are used for jobs which can be efficiently solved by dividing tasks into parallel subtasks being processed by different agents. Agents with different goals, even if competitive, can interact in a common environment. Agents can be i.e. software programs. The following aspects must be considered:

- The set of actions
- Knowledge and information about self and environment
- Interaction, i.e. coordination, cooperation and communication between individuals
- Modelling of perception, reasoning, behaviour and planning
- Implementation and technical restrictions

As a famous example the online dictionary Wikipedia can be seen as a Multi-Agent-System. The set of actions consists of manipulating, creating and organizing pages. The set of agents consists of registered and unregistered users, administrative users and software agents. All agents are connected by various types of message passing. For unregistered users, communication is almost unidirectional. Implementation and technical restrictions are based on the HTTP protocol.

2 The Multi-Agent-System Simulation (Massim) Project

The aim of the Massim Project[3] was to create an executable implementation of MAS for permanent changing environments. It was developed as a framework for the inter-

[1]E-mail: slawomir.deren@tu-clausthal.de
[2]Department of Informatics, Clausthal University of Technology, Clausthal-Zellerfeld, Germany
[3]http://agentmaster.in.tu-clausthal.de

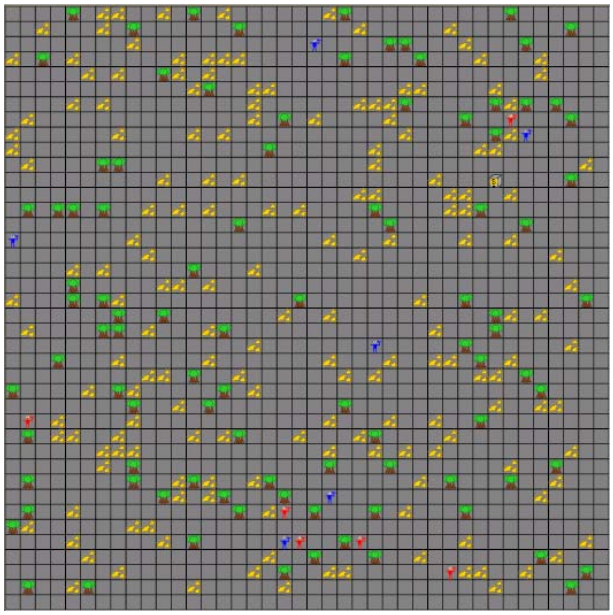

Figure 4: The playing field

national CLIMA Contest[1]. The scenario of the 2007 contest was based on the scenario of CLIMA VII Contest 2006.

In general the participating agent teams are executed on local hardware while the simulated environment runs on the contest simulation server. The communication protocol is based on XML and all agents have to connect with password and username via Internet. The server provides informations to all agent clients, and agents response with their actions. If no action is transmitted within five seconds, a timeout fires and the server returns an invalid action to the environment. The server is adaptive with respect to the simulations and it is possible to create arbitrary environments. In this contest the environment was a rectangular grid consisting of cells as shown in the figure 4. Each cell can contain one of the following objects: Agent, depot, gold, obstacle and mark (string with maximal 5 characters). An agent is not able to occupy a cell containing another agent or obstacle.

In this contest each team consisted of 6 agents and the goal was to collect as much gold items as possible. The server sent a message in XML format to provide the necessary information for performing actions in each simulation step. All agents should simulate locally the environment and response with an action. Follow actions are possible:

[1] http://cig.in.tu-clausthal.de/CLIMAContest/

- Up, left, down, right
- Pick, drop
- Mark, unmark
- Skip

The simulation engine had to compile all actions and had to create a new state of the environment. For obtaining this goal we created the follow algorithm:

1. Filtering of impossible actions
2. Computing of probability of action failure:

   Each agent's action can fail. The simulation engine computes the probability by the following formula:

   $$P = P_{sim} + \frac{P_{max} - P_{sim}}{N_{\mathrm{ItMax}}} \cdot N_{\mathrm{It}}$$

   where P_{sim} = default value of probability of action failure,

   P_{max} = maximal value of probability of action failure,

   N_{ItMax} = maximal amount of carried gold items and

   N_{It} = current amount of carried gold items
3. Detecting of possible collisions between agents on the grid
4. Processing of pushing. Only one agent is able to push in one simulation step
5. Replacing of not yet performed move actions with skip action
6. Executing of all no move actions

After that other internal changes like generation of gold items on the grid are executed.

3 Contest 2007

The contest started on Wednesday, 2nd May 2007 and ended Sunday, 6th May 2007. We organized all matches between participating teams. Following teams participated:

- *JACKteam* and *GOLOGteam* from RMIT University, Australia
- *microJiacteam* and *JiacIVteam* from TU Berlin, Germany
- *FLUXteam* from TU Dresden, Germany

- *APLteam* from University of Utrecht, Netherlands
- *Jasonteam* from University of Durham, United Kingdom and Universidade Regional de Blumenau, Brazil
- *AC07bot* from TU Clausthal, Germany for benchmark and testing

Pos.	Team	Gold score	Diff.	Points
1	JiaclVteam	2824 : 1759	1065	63
2	microJiacteam	2680 : 1598	1082	54
3	Jasonteam	2563 : 1988	575	49
4	FLUXteam	2514 : 1816	698	43
5	APLteam	1246 : 2585	-1339	12
6	JACKteam	730 : 2811	-2081	3

Table 1: Contest results

The tournament consisted of matches between participants. A match is a sequence of simulations in which two teams competed in five different scenarios. We called them "park", "fence", "semiramis", "meadow" and "overkill". Each scenario had a different difficulty factor. For example "fence" was a grid with the following settings:

- size 51x51 cells
- 1000 steps
- depot at the cell [29, 34]
- 235 obstacles and 155 gold items on the grid

During the tournament the team GOLOG was replaced by a dummy agent because of technical problems. Table 1 shows the end result of the contest (without GOLOG team and AC07 bot).

References

[1] M. Dastani, J. Dix, P. Novak. Multi-Agent Programming Contest 2007. *http://cig.in.tu-clausthal.de/AgentContest2007/*, 2007.

[2] M. Dastani, J. Dix, P. Novak. The Second Contest on Multi-Agent Systems based on Computational Logic. *IfI Technical Report Series*, IfI-06-13, 2006.

An evolutionary approach to Tetris

Patrick Dohrmann[1,2,3]

1 Introduction

This article is based on the paper 'An Evolutionary Approach to Tetris' by Niko Böhm, Gabriella Kókai and Stefan Mandl, MIC 2005. Tetris has been invented in the 80s and is a very popular game. The goal addressed in the article is to learn to play the Game of Tetris using genetic algorithms (GA). Tetris is NP-hard even if the sequence of tetraminos is known in advance, as proved by Demaine. Therefore, heuristic approaches such as GA have to be used.

2 Tetris

Tetris was invented in 1985 by Alexei Paschitnow and is still a popular game with many clones. The game-board consists of 10 by 20 grid cells where seven different tetraminos fall downwards. Each piece occupies four grid cells and has a different shape: L-shape, J-shape, S-shape, Z-shape, T-shape, I-shape and O-shape. In every step, a random piece enters the top of the board. The player can move it horizontally and rotate it while it is falling downwards. The tetramino stops if it hits the ground or another tetramino. If the player fills one or up to four lines at the same time, these lines are removed from the board and a score is obtained. The game ends when no tetramino is able to enter the board. There is no way to win the game; the goal is to obtain a score as high as possible.

3 Genetic Algorithm

A genetic algorithm (GA) is a heuristic for optimizing problems which are NP-hard or have a big search space. It is based on the evolution of individuals and traces back to

[1]E-mail: patrick.dohrmann@tu-clausthal.de
[2]Department of Informatics, Clausthal University of Technology, Clausthal-Zellerfeld, Germany
[3]Please note that this abstract does not contain original work but it summarizes an algorithmic apporach presented by Böhm et al. as indicated in the reference.

Charles Darwin: *Survival of the fittest*. Each individual consists of a genotype and a phenotype. The genotype holds the information of an individual and it describes the phenotype which declares the behaviour and appearance of it. The genotype can be represented as a gene string which is once initialized at random. Then the fitness of every individual of the current population is evaluated and good individuals are selected for further evolution. There are many ways to select good individuals, e.g: fitness-proportionate selection (the expected number of children scales with the fitness value); tournament selection (of two randomly selected individuals, the better one will survive with higher probability); rank based; roulette-wheel and many others. After selecting the individuals, they recombine their information to create a new generation. Often a one- or two-point crossover is used, i.e., one or two crossover points are selected randomly and the parts in between are interchanged. After crossover, the positions of each individual are randomly mutated using e.g. bit-flips. This process, fitness evaluation, selection, crossover, and mutation, is repeated until a satisfactory population has evolved.

4 Learning Tetris

To apply genetic algorithms to Tetris, the genotype and phenotype have to be defined: the strategy and its representation. When a new tetramino enters the game board, a strategy has to find a good move, i.e. it has to decide the rotation and placement of the tetramino. For this purpose, it is sufficient to find a rating function which evaluates every possible move and then choose the best one. This rating function can be defined as a combination of simpler functions which extract important features from the board and which are combined to yield an overall evaluation. Such features can include the maximum pile height, the well depth, the number of cleared lines, the amount of holes, etc. The following three combinations of the elementary features $r_i(b)$, b being the board, are chosen:

- linear rating function: $R_l^*(b) = \sum_{i=1}^{n} \omega_i \cdot r_i(b)$
- exponential rating function: $R_e^*(b) = \sum_{i=1}^{n} \omega_i \cdot r_i(b)^{e_i}$
- exp. rating function with displacements: $R_d^*(b) = \sum_{i=1}^{n} \omega_i \cdot |r_i(b) - d_i|^{e_i}$

Thereby, w_i, d_i, and e_i are numbers which have to be evolved by the GA, i.e. the weight-chromosome $\omega = (\omega_1 \dots \omega_n)$, the exponents-chromosome $e = (e_1 \dots e_n)$, and the displacement-chromosome $d = (d_1 \dots d_n)$.

The fitness value for each individual of the population is chosen as follows: every individual is evaluated in a number of exemplary games. This guarantees, that higher fitness values correspond to a better performance of the game. The exact values are the amount of placed tetraminos, the number of cleared lines, or the number of occupied cells, three equivalent measures for a game an individual plays. The fitness value is

now computed as the arithmetic mean of a fixed number of played games of an individual. The remaining GA consists of standard fitness-proportional selection, two-point crossover, and mutation using Gaussian noise. So the whole algorithm is as follows:

```
while (unsatisfied) {
   for all individuals {
      for all repetitions i {
         play game_i
         perf_i = measure of game_i }
      fitness_i = sum of perf_j/ number of games }
   selection
   recombination
   mutation }
```

5 Results

The genetic algorithm was run with a population of 100 individuals on a 10 by 20 game-board. The linear rating function managed to clear about 170000 lines and terminated after 20 to 30 generations. The exponential rating function took three months to terminate after the 30th generation but succeeded in clearing about 5 million lines. The search space in this case is very exhaustive because of two parts of the chromosomes and a long running time of the games. Therefore the next runs were tested on a smaller game-board (6 by 12), which can give some hint on the overall behavior of strategies, although not necessarily exact results for the larger board. In particular, there seems to be a tendency that linear rating performs better on a larger board than exponential rating.

After learning the weights, a strategy for playing Tetris is available. This can be analyzed: It turns out that keeping the piles low, avoiding holes, and keeping the surface of all placed pieces plain are the most important criteria. This is quite similar to a human player. The other features were rated with smaller weights depending on the solution at hand.

6 Conclusion

The results of the implemented GA are similar to other results reported in the literature which are based on GAs or alternative methods such as reinforcement learning. Adding more criteria for the rating of the game boards might further increase the performance due to larger expressiveness but it also severely extends the search space.

References

[1] N. Böhm, G. Kókai and S. Mandl. An Evolutionary Approach to Tetris. *MIC2005: The Sixth Metaheuristics International Conference.* Vienna, Austria, August 22–26, 2005.

Hyper-heuristics and evolutionary computing algorithms for technicians and interventions scheduling

Pawel Lichocki*[1,2]*, Grzegorz Pawlak*[2]*, Slawomir Bak*[2]*,

***Wojciech Mruczkiewicz*[2]**

1 Introduction

The problem of scheduling technicians and interventions was formulated for Challenge 2007 ROADAF [1]. There is given a set I of n interventions and the set T of m technicians and unlimited number of days. The technicians group into teams and perform the interventions. Each intervention has certain requirements (in many domains and on many levels), which must be fulfilled by a team performing it. The goal is to schedule interventions and technicians in such a way that total cost of the schedule is minimized. The cost of the schedule is given by the formula

$$28t_1 + 14t_2 + 4t_3 + t_4$$

where t_1, t_2, t_3 and t_4 are appropriately ending times of the last scheduled interventions of priority 1, 2, 3 and the ending time of the last intervention regardless the priority. There are also other constraints which make this problem particularly difficult to solve - as for example precedence constraints of interventions, unavailability of technicians on different days, the possibility to abandon (outsource) the intervention for a specific amount of virtual money.

The goal of this research was to construct a heuristic algorithm solving the above described scheduling problem. The first step was to design a simple, greedy dedicated heuristic. Next, it was used to create a hyper-heuristic scheme in which it was combined with an evolutionary algorithm. Finally, it was shown that the second hierarchical approach is superior to a constructive heuristic.

[1]E-mail: plichocki@skno.cs.put.poznan.pl

[2]Institute of Computing Science, Poznan University of Technology, Poznan, Poland

2 Dedicated heuristic scheme

```
init();
while (!pool.empty())
    intervention = pool.getNextCandidate();
    if (null != intervention)
        schedule.tryAdd(intervention);
    else
        schedule.createNewDay();
        pool.reset();
schedule.improve();
```

The idea of interventions pool was introduced. It is an iterator from which (basing on different measures, as for example interventions' priorities) each intervention candidate is pulled. Than, the algorithm tries to schedule the intervention candidate on the currently processed day. If this operation is successful the scheduling of interventions still continues at the same day. If it fails the pool is reset and new empty day is added to the schedule. After creating entire schedule the solution is improved by rescheduling and swapping interventions if it leads to bettering the global solution.

This is a greedy like heuristic. The schedule created by it has a good quality at the beginning and very poor at the end. Therefore, there is a need to apply more sophisticated methods which will allow the algorithm to take into account the optimality of the entire schedule, not only of the sub-schedules for each day separately.

3 Hyper-heuristic scheme

The abstract space of permutations has been defined. In this space the search was performed by the evolutionary computing according to the following scheme [2]:

```
init();
while (!end)
    select();
    breed();
    mutate();
```

The most important part is the selection, where individuals' evaluation must be done. To do this previously implemented simple, greedy heuristic was used. But in this case the process of pulling the interventions out of the pool was using the given permutation to determine the order (instead of arbitrary measures applied previously).

It might be stated that the evolutionary heuristic makes a use of the greedy one to map the permutations' space into solutions' (schedules') space. This hierarchical approach is called a hyper-heuristic method.

4 Results and conclusions

It was found out that hyper-heuristic greatly improves the quality of the results. In 19 on 20 instances it led to bettering the global solution. However, for one instance this approach resulted in a worse solution. Therefore, several steps for the future might be defined as follows:

1. one may try to change the meta-heuristic searching engine to other than evolutionary computing (for example to tabu search)
2. one may try to change the dedicated heuristic which maps permutations into solutions (for example to dynamic programming).

All of this will lead to a general hyper-heuristic framework, which will allow mapping between permutations' and schedules' spaces. It is a very promising method, because it appears that it is possible to apply this scheme to solve other combinatorial problems which are easily mapped to a permutations' space.

References

[1] P.F. Dutot, A. Laugier and A.M. Bustos. Technicians and Interventions Scheduling for Telecommunications. *http://gilco.inpg.fr/ChallengeROADEF2007/en/sujet/sujet2.pdf, France Telecom R&D*, 2006.

[2] D. Goldberg Genetic Algorithms in Search. *Optimization and Machine Learning, Addison-Wesley*, 1989.

TU Clausthal
ICOLE 2007, LESSACH, AUSTRIA

Polynomial time algorithm for coupled tasks scheduling problem

Michal Tanas[1,2]***, Jacek Blazewicz***[3]***, Klaus Ecker***[4]

1 Introduction

One branch of scheduling theory is concerned with scheduling of coupled tasks. A task is called *coupled* if it contains two operations where the second has to be processed some time after a completion of the first one. This variant of scheduling problem often appears in radar-like devices, where two radar pulses are used to calculate trajectory and speed of a moving plane.

The complexity of various scheduling problems with coupled tasks has been studied in [3]. Although most of the cases are NP-hard [3], some polynomial algorithms were found in [4].

It is proven in [2] and in [5] that problem of scheduling of strictly precedence related coupled tasks where the criterion is to minimize schedule length is NP-hard in the strong sense if the precedence constraints have form of a general graph even if there is only a single machine in system, all gap lengths are equal and all precesing times are equal to 1.

In this presentation, we complement the above result by presenting the proof of polynomial time solveability of the problem stated above under assumption that the precedence constraints graph has a form of chains.

2 Problem formulation

We consider the problem of scheduling n coupled tasks on a single machine, where each coupled tasks T_i consists of two operations T_{i1} and T_{i2} and a *gap* between them.

[1]E-mail: Michal.Tanas@cs.put.poznan.pl
[2]Applied Computer Science Division, Physics Faculty, Adam Mickiewicz University, Poznan, Poland
[3]Institute of Computing Science, Poznan University of Technology, Poznan, Poland
[4]Center for Intelligent, Distributed and Dependable Systems, Ohio University, Athens, USA

During the gap, another task can be processed. Let p_{i1} and p_{i2} denote the processing times of operations T_{i1} and T_{i2}, respectively.

The gap is exact when operation T_{i2} has to start exactly l_i units of time after the end of operation T_{i1}, where l_i denotes a length of the gap. In this paper, the only cases considered are those where all l_i are equal, i.e. $l_i = l$, $i = 1, 2, \ldots, n$.

Precedence constraints of coupled tasks can be strict or weak. $T_i \prec T_j$ means that $T_{i2} \prec T_{j1}$ if precedence constraints are *strict*, and $T_{i2} \prec T_{j1} \ \wedge \ T_{i2} \prec T_{j2}$ if they are *weak*.

The special case of a coupled task problem involves identical tasks. Commonly, such tasks are denoted by (p_1, l, p_2), where $p_1 = p_{i1}$, $p_2 = p_{i2}$, $l = l_i$ for all $1 \leq i \leq n$.

Adapting the commonly accepted notation for scheduling problems the scheduling problem considered in this paper can be denoted by $1|(1, l = const, 1) - coupled$, $strict\ chains$, $exact\ gap|C_{max}$, which means that there is one processor in a system, tasks are coupled and identical with processing times $p_{i1} = p_{i2} = 1$, $\forall_{1 \leq i \leq n}$, gaps are exact and have uniform length l, precedence constrains are strict and the optimization criterion is to minimize the schedule length $C_{max} = \max\{t_{j2}\}$, where t_{j2} is a completion time of T_j (its second operation).

3 A polynomial time algorithm

To create an optimal schedule for a given instance of the problem $1|(1, l = const, 1) - coupled, strict\ chains, exact\ gap|C_{max}$ two things should be found — the optimal order of tasks and in what time units of the schedule the machine should remain idle.

To find the optimal order of tasks it is enough to convert the given instance of the problem $1|(1, l = const, 1) - coupled, strict\ chains, exact\ gap|C_{max}$ into the corresponding instance of the problem $P(l + 1)|pmtn|C_{max}$, then solve this instance using a McNaughton-like algorithm and the optimal order of coupled tasks can be discovered from the gained parallel-system schedule.

The idea of the second stage is based on observation that any feasible schedule for the problem $1|(1, l = const, 1) - coupled, strict\ chains, exact\ gap|C_{max}$ can be decomposed into a sequence of partial schedules (called *segments*) each of them contains an upper-bounded number of pairwise independent coupled tasks and the length of each such segment is also upper-bounded. This observation limits the solution space to a polynomial size and thus allows us to exploit an exhaustive search of the solution space which in this case remains polynomial.

4 Conclusion

The considered problem $1|(1, l = const, 1) - coupled, strict\ chains, exact\ gap|C_{max}$ can be solved in polynomial time.

References

[1] K. Baker, *Introduction to Sequencing and Scheduling*, J. Wiley, New York, 1974

[2] J. Blazewicz, K. Ecker, T. Kis, M. Tanas Complexity of Scheduling Coupled Tasks on a Single Processor *Journal of Brazilian Computer Society* **N 3 vol 7**, 2002, 23-26

[3] A. J. Orman, C. N. Potts, On the complexity of coupled tasks scheduling, *Discrete Applied Mathematics* **72**, 1997, 141-154.

[4] A. J. Orman, C. N. Potts, A. K. Shahani, A. R. Moore, Scheduling for the control of a multifunctional radar system, *European Journal of Operational Research* **90**, 1996, 13-25.

[5] M. Tanas *Scheduling of Coupled Tasks* Papierflieger, Clausthal-Zellerfeld, 2004

Minimization the time interval on the car assembly line

Grzegorz Pawlak[2]***, Tomasz Kujawa***[1,2]

1 Introduction

In this paper the problem of minimization the time interval in the assembly line was considered. The motivation for the research was drown from the real car factory. The original car assembling problem is formulated as a permutation flow shop problem in the multi-stage system [1] because the production is synchronized by the time period so the scheduling problem could be modeled as a scheduling problem on the parallel machines with machines eligibility. Generally, the operations assigned to the stage should be processed during the particular time period. The assembly line is flexible in a sense that some assembly operations could be done alternatively on the one out of the several stages. The purpose is to minimize the time interval which is equivalent to minimize the flow time. The above described problem is NP-hard. In the literature the balancing of the assembly line was considered for example in [3].

Actually, there are different car models and cars are equipped differently. The mathematical model of above problem has been created and algorithms has been constructed.

2 Problem formulation

Each car has certain requirements, which must be fulfilled during the assembly process. For each type of car the directed graph of the precedence constraints is also defined. For each operation the subset of stages is defined. There is a sequence of cars which should be assembly.

The goal is to assign the operations to stages (in the model the parallel processors) in such a way that the time interval will be minimized respecting the precedence constraints.

[1]E-mail: tkujawa@skno.cs.put.poznan.pl
[2]Institute of Computing Science, Poznan University of Technology, Poznan, Poland

3 Solution algorithms

For this problem the branch and bound algorithm was proposed. The objective of this algorithm is to find the shortest time interval and check if all types of cars can be assembled during that time interval. The idea of the algorithm is to check all possible topological orders of the operations and stages combinations for each car type.

Main points of the algorithm are as follows:

1. Initial solution - Greedy like heuristic choosing the first type of the car.
2. Algorithm *A1* - Find the minimum interval of that type.
3. Algorithm *A2* - Execute the B&B algorithm for each type of the car from the potential solutions.
4. If list of types is not empty go to the point 1.

Initial solution The initial type of the car is chosen in the greedy way from the list of potential solutions. This algorithm finds the car type which has the longest total processing time of operations.

Algorithm A1 For chosen car type this algorithm calculate the shortest time interval T^*. As a result it returns the sequence of operations and the time interval for that sequence.

Algorithm A2 This algorithm is checking the rest car types from the list of the potential solutions if they can be done within that time interval T^*. If the type can be done in the time interval T^* this type can be deleted from the list.

4 Conclusions

The problem of finding the shortest time interval was defined and the branch and bound algorithm was proposed. The future research will be concentrated on the construction of metaheuristic algorithms which can base on the above method. Other methods are also proposed in [2] and these algorithms should be compared to the presented algorithm. Future research will also dealt with the workers teams assignment to the stages.

References

[1] Escudero L., Sciomachen A., The Job Sequencing Ordering Problem on a Card Assembly Line, Optimization in Industry, Wiley and Sons, New York 1993, pp. 251-262.

[2] Puchta M., Gottlib J., Solving Car Sequencing Problems by Local Optimization., Applications of Evolutionary Computing, LNCS 2279, Springer,2002, pp. 132-142.

[3] Sawik T.,, Production Planning and Scheduling in Flexible Assembly Systems, Springer-Verlag, 1999.

TU Clausthal
ICOLE 2007, LESSACH, AUSTRIA

Practical scheduling problems in the car factory

Grzegorz Pawlak[1,2]

1 Introduction

In the paper the production scheduling problems drown from the car factory were considered. The car production system is divided into three general fairly independent main shops namely body shop, paint shop and assembly line. In the many places in the production process the scheduling problems arise. The general purpose is to optimise the production system it means to increase the productivity. In the paper the revision of some practical scheduling problems in the car factory was presented. There was developed several scheduling models for the specific problems met in these three production stages. Generally, the goal of the production system is to increase the efficiency of the production and increase the car flow that is equivalent to the maximization of the throughput rate. Each shop can be considered separately with the different goal functions. The optimisation methods for the Renault car factory were presented in [2]. For this some problem one of approaches were proposed in [1]. Many other authors were tackling the car scheduling problems for example [3].

2 Body shop scheduling problems

The scheduling problems in the body shop had the different particular goal function and in opposite to the paint shop it has the different gals as in the assembly line. At the first step the production process has been analysed and the scheduling models were contracted. The specific goal functions for each stage has been formulated. In the body shop the robotic flow lines ware analyse from two stand points. First, there were the assumption that the production is synchronous then the minimum time period for each robotic cell was calculated. Thus the scheduling model is equivalent to the scheduling

[1]E-mail: grzegorz.pawlak@cs.put.poznan.pl
[2]Institute of Computing Science, Poznan University of Technology, Poznan, Poland

task on the parallel machines with the machines eligibility. The same situation appears in the assembly line where the assembly operations are processed manually by the worker teams. In this case the same scheduling model can be used for the determination of the time interval at these stages. In case of ideal synchronous production there is the fix time interval for each either robotic cell or assembly stage. In this case one can found the minimum value of such period for each type of produced cars. The scheduling problem on eligible parallel machines is NP-hard and when the precedence constraints were introduced (as it is in the practical case) then the problem becomes strongly NP-hard. As the solution the branch and bound algorithms were proposed and tested also with constructed heuristic algorithms. On the other hand the asynchronous production the system become the classical multi-stage flow shop with additional constraints.

3 Paint shop scheduling problems

The goal function for the paint shop is to have the car sequence where the number of the colour changes is minimum. There was formulated the objective function minimizing the length of the schedule for the given time horizon which is equivalent to the minimum colour changes. The on-line algorithm for the sorting buffer management was proposed, implemented and computational experiment performed. It was shown, for real historic data that there is significant improvement in comparison with the previously implemented algorithms. The model of the re-entrant jobs at repair line for the car bodies which need more than once painting process performed (because of painting quality reasons) was proposed and analysed. Also the measurement of the differences between the car sequences were defined. This measurement function was used to predict incoming car sequence and to use this information for building more sufficient buffer management system due to the minimum colour changes objectives.

4 Assembly line scheduling problems

For the assembly line the main goal is to prepare the *smooth* car sequence from the assembly point of view. It means that the best it is to have the balanced workload for each assembly stage. To estimate the quality of the car sequence from that point of view the two coefficient functions were defined and normalized to the range from 0 to 1. Then, one can evaluate the given car sequence and calculate the *assembly difficulty* according to the balance coefficients. Similar to body shop the model of synchronous production can be used the same scheduling model to define the minimum time period for each stage of the assembly line. In case of asynchronous assembly lines the model with the processing time restriction for the operations were proposed and the solution algorithms branch and bound, meta-heuristics local search and tabu search were constructed. The computational experiments have been performed.

5 Conclusions

Form many practical production problems the scheduling models have been constructed and solution algorithms have been proposed. Mostly, the branch and bound methods were developed and then the solution compared with the heuristic algorithms to estimate the latter quality. There were proposed the meta-heuristic approach local search, tabu search and genetic algorithms. In many cases the constructive heuristic were proposed too. Computational experiments comparing the quality and showing the efficiency of proposed algorithms were performed. In most cases the instances for computational experiments were taken from the Volkswagen car factory located in Poznan. The one of the version of on-line algorithm was implemented in that factory. There are may direction for the further research, especially in the assembly line. There is the problem of overlapping time intervals (the workers' teams not always keep the time interval restriction). Also the problem forming workers' teams matching them to the stages with specified operations is the practically important and will be analysed in the future.

References

[1] G. Pawlak, P. Piechowiak, M. Plaza and M. Rucinski, Local Search Algorithm for the Car Sequencing Problem, Conference Proceedings, *6eme congres de la Societe Francaise de Recherche Operationnelle et d'Aide a la Decision, Tours.* 14-16.02.2005, p. 37–38

[2] Ch. Solnon and V-D. Cung and A. Nguyen and Ch. Artigues, The car sequencing problem: overview of state-of-the-art methods and industrial case-study of the ROADEF'2005 challenge problem. *European Journal of Operational Research*, forthcoming volume, 2007.

[3] T. Warwick and E. Tsang. Tackling car sequencing problems using a genetic algorithm *Evolutionary Computation*, vol. 3., No. 3., 1995, pp. 267–298